ON THE LIMITATIONS OF LLM-SYNTHESIZED SOCIAL MEDIA MISINFORMATION MODERATION

Sahajpreet Singh, Jiaying Wu, Svetlana Churina, Kokil Jaidka
National University of Singapore
sahajpreet.singh@u.nus.edu, {jiayingw,churinas,jaidka}@nus.edu.sg

ABSTRACT

Despite significant advances in Large Language Models (LLMs), their effectiveness in *social media misinformation moderation* – specifically in generating high-quality moderation texts with accuracy, coherence, and citation reliability comparable to human efforts like Community Notes (CNs) on X – remains an open question. In this work, we introduce MODBENCH, a real-world misinformation moderation benchmark consisting of tweets flagged as misleading alongside their corresponding human-written CNs. We evaluate representative open- and closed-source LLMs on MODBENCH, prompting them to generate CN-style moderation notes with access to human-written CN demonstrations and relevant web-sourced references utilized by CN creators. Our findings reveal persistent and significant flaws in LLM-generated moderation notes, signaling the continued necessity of incorporating trustworthy human-written information to ensure accurate and reliable misinformation moderation.

1 INTRODUCTION

Moderating misinformation on social media is crucial for effective governance and intervention in online discourse (Lazer et al., 2018). Moderation texts play a key role in helping information consumers make more informed decisions by countering misinformation with trustworthy, fact-based evidence. A prominent real-world example is Community Notes [1] (CNs), a feature on X where users collaboratively generate moderation notes that provide factual explanations along with cited sources (Chuai et al., 2024).

Despite these collaborative efforts, maintaining the timeliness and coverage of CNs remains labor-intensive, requiring significant human intervention. To address this challenge and scale misinformation moderation, open-source large language models (LLMs), such as LLaMA-3 (Dubey et al., 2024) and Qwen-2.5 (Yang et al., 2024), have emerged as promising alternatives due to their strong capabilities in generating misinformation-related explanations (Hu et al., 2024; Qi et al., 2024; Zhou et al., 2024). However, the extent to which LLM-generated moderation notes can effectively replace human-written CNs remains an open question.

LLMs, trained on vast web data, contain abundant internal knowledge about facts. Combining this with their strong reasoning capabilities, LLMs can generate moderation notes under the CN format. In this paper, we investigate the core research question: **Can LLMs synthesize CNs that match the quality of human-written ones?** To this end, we introduce MODBENCH, a real-world misinformation moderation benchmark consisting of misleading tweets (posted before LLM knowledge cut-off date) and their corresponding human-authored CNs. In line with the CN format, we synthesize moderation notes with diverse LLMs – each note comprising a factual explanation and supporting sources – and systematically evaluate their quality.

While LLM-generated notes are often relevant, our findings reveal two fundamental limitations. (1) Source invalidity: many generated links are non-existent or unverifiable. (2) Structural inconsistencies: failure to consistently adhere to the instructed CN format. Our benchmark and findings offer valuable insights for future research on automated misinformation moderation and the evolution of

[1] https://communitynotes.x.com/guide/en

CNs. As current LLM-based moderation consistently falls short, our results highlight the necessity of integrating human-authored evidence into LLMs to enhance accuracy and reliability.

2 MATERIALS & METHODS

2.1 MODBENCH DATA CURATION

Community Notes (CNs) play a crucial role in combating misinformation on X/Twitter by providing contextualized explanations supported by reliable references. In this study, we investigate the limitations of automated moderation by constructing an English tweet-notes benchmark. Our dataset comprises a large collection of English-language tweets paired with their corresponding CNs, which provide additional information/context about potentially misleading posts. Specifically, we consider tweet-note pairs where the tweets are labeled as "MISINFORMED_OR_POTENTIALLY_MISLEADING" and the notes are rated as "HELPFUL". The dataset spans from 2018 to 2023 and, after processing, contains a total of 133,436 entries.

2.2 EXPLORING LLMS FOR AUTOMATED CN GENERATION

We explored the potential of leveraging LLMs for automated online content moderation by generating CNs through in-context learning, specifically (1) **few-shot learning** (Brown et al., 2020). Initially, we employed few-shot prompting, instructing the LLM to generate a CN for a given misleading tweet. To further assess its capability, we (2) **provided relevant factual information** to determine whether LLMs, trained on past data, could accurately extract and cite key details when given a web link as a reference. In our final experiment, we incorporated additional context related to CN composition, simulating the information available to a human CN writer. We further employed the (3) **Chain-of-Thought (CoT)** (Wei et al., 2022) approach to enhance reasoning and factual grounding. Prompt structures used in these experiments are detailed in Appendix A.

2.3 EXPERIMENTAL DETAILS

We conducted our experiments on two open-source models: (1) **Llama-3.1-8B-Instruct**[2] and (2) **Qwen-2.5-7B-Instruct-1M**[3]; along with one representative closed-source model, (3) **GPT-4o-mini-2024-07-18**[4]. For computational resources, we utilized an H100 GPU with float16 precision to load and run the models efficiently.

As this study serves as a proof of concept, we conducted experiments on a randomly sampled subset of 1,000 tweets, ensuring reproducibility by setting a fixed random seed of 13. For the few-shot learning setup, we selected four different examples, chosen based on random state of `tweetID% 13`, to maintain both reproducibility and sufficient variation in prompts for in-context learning.

To evaluate LLM-generated CNs, we employed standard automated text-generation metrics, including **BLEU** (Papineni et al., 2002), **METEOR** (Banerjee & Lavie, 2005), **ROUGE-L** (Lin, 2004), and **BERTScore** (Zhang et al., 2019). Additionally, we conducted human assessments focusing on specificity, usefulness, truthfulness, and reference validity to provide qualitative insights into the generated moderation notes.

3 EVALUATION AND ANALYSIS

3.1 QUANTITATIVE EVALUATION

From Table 1, we make the following key observations. **(1)** Few-shot without web references yields the weakest results, which aligns with expectations, as LLMs often struggle with detailed or less widely known facts when relying solely on internal knowledge. **(2)** Few-shot with web references outperforms few-shot CoT with web references, which suggests that direct access to factual sources

[2] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`
[3] `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M`
[4] `https://platform.openai.com/docs/models`

Table 1: Results on automated metrics (* represents closed-source model).

Model	Method	BLEU	METEOR	ROUGE_L	BERTScore
Llama-3.1	Few-shot	0.046	0.191	0.108	0.810
	+ Web References	0.156	0.264	0.239	0.863
	+ CoT	0.146	0.255	0.224	0.863
Qwen-2.5	Few-shot	0.044	0.193	0.108	0.811
	+ Web References	**0.179**	**0.287**	0.263	0.869
	+ CoT	0.104	0.246	0.193	0.846
GPT-4o-mini*	Few-shot + Web References	0.174	0.284	**0.277**	**0.871**

is more impactful than additional reasoning steps in improving moderation quality. **(3)** Surprisingly, GPT-4o-mini shows only marginal improvement over smaller open-source LLMs. Despite its extensive world knowledge and advanced capabilities, GPT-4o-mini performs only slightly better than its smaller open-source counterparts, leading us to remark, *"We can't believe it's not much better"*.

Our proof-of-concept highlights an important insight: **smaller open-source models, when provided with direct access to relevant factual resources such as web links, can achieve comparable performance to larger proprietary models**. These findings highlight the potential for more accessible, transparent, and cost-effective solutions empowered by open-source LLMs in automated misinformation moderation.

3.2 Qualitative Inspection

To evaluate the performance of Qwen-2.5 and GPT-4o-mini when prompted using few-shot learning with web links, we inspected their outputs based on specificity, usefulness, truthfulness, and reference validity. We make the following observations:

(1) Both models generally incorporated the provided web references. However, GPT-4o-mini occasionally omitted references when multiple sources were available, likely due to output length constraints. In contrast, Qwen-2.5 sometimes generated incomplete or malformed links, leading to citation errors. For example:

> **Web Reference:** `https://www.timesofisrael.com/amid-ongoing-war-bp-and-eni-among-firms-awarded-gas-exploration-licenses-in-israel/`
> **Incomplete Link Generated by Qwen-2.5:** `https://www.reuters.com/business/energy/israel-awards-gas-exploration-`

Incomplete references like these undermine the reliability of the generated notes, as users may be unable to verify the claims, reducing their overall trustworthiness.

(2) Structural inconsistencies in smaller, open-source Qwen-2.5. In addition to citation errors, Qwen-2.5 exhibited formatting inconsistencies. Rather than generating a well-structured response, it often introduced factual information first, followed by an unnecessary "Community Notes:" section, despite no explicit formatting requirement. Furthermore, it occasionally retained prompt artifacts or referred to itself as an AI assistant, making its output appear less natural and more AI-generated compared to human-written notes. For example, in the example below, the model appends unnecessary third-party comments to the generated CN, reducing overall readability and coherence.

> On November 28, 2023, Ecuador's National Police arrested a suspect in relation to the murder of a presidential candidate. The suspect was later found dead under suspicious circumstances. No evidence of hitmen or a military enclave being involved has been reported. This incident is still under investigation. https://twitter.com/FiscaliaEcuador/status/
> The community note provides factual context about the ongoing investigation into the murder of a presidential candidate in Ecuador, dispelling claims of hitmen and a military enclave involvement. It references the official statement from the Fiscala de la Nacin Ecuatoriana, which is the national prosecutor's office in Ecuador, indicating that the case is under investigation and no conclusive evidence of the additional claims has been reported.

(3) GPT-4o-mini demonstrated stronger structure and coherence. Compared to Qwen-2.5, GPT-4o-mini produced more structured and coherent responses, aligning better with the original tweet and focusing on explaining why the content might be misleading or inaccurate while providing supporting evidence. For example:

> The claim about a presidential candidate being murdered under police custody is misleading. While there are concerning incidents involving hitmen, the official investigation is ongoing, and there is currently no verified evidence indicating that any murder took place under police custody or that these events are connected to framing left-leaning politicians. For accurate updates, refer to the official statements from the Ecuadorian authorities.

In contrast, Qwen-2.5 prioritized factual accuracy but lacked contextual engagement with the tweet, making its responses dense and harder to process in fast-paced social media environments. While it often included accurate information, it failed to clearly connect its explanations to the misinformation at hand, reducing its effectiveness for moderation.

(4) Both models successfully cross-checked information from provided sources. However, Qwen-2.5 occasionally fabricated references when no web links were given, presenting them as factual. GPT-4o-mini, while also leveraging internal knowledge, did not fabricate links; instead, it structured responses around existing verifiable information.

Overall, qualitative investigation demonstrated benefits of GPT-4o-mini over Qwen-2.5 due to its clearer structure, better alignment with the original tweet, and more readable explanations. While Qwen-2.5 sometimes provided more detailed responses, its formatting inconsistencies, incomplete references, and tendency to generate artificial citations made it less reliable for fact verification.

3.3 Implication for Future CN Research

Our observations, based on both automatic evaluation scores and human assessment, indicate promising potential for developing a small open-model-based automated CN generator – provided that the models are given access to relevant factual information (Bommasani et al., 2021). While we identified some shortcomings in these smaller open models, their performance remains impressive, especially considering that no fine-tuning was performed for formatting. Moreover, many of the structural errors appear fixable, suggesting that improvements could be made with reasonable effort.

Another study, HelloFresh (Franzmeyer et al., 2024), explores the use of LLMs and web data to assess the usefulness of community-driven factual information, such as X's Community Notes and Wikipedia edits. In the future, HelloFresh could serve as an evaluation metric for automated CN generation tasks.

For deeper insights, future work could explore providing factual references as textual inputs rather than as hyperlinks. Additionally, incorporating web-based agents (Deng et al., 2024) to automate CN generation appears to be a promising direction. However, we emphasize that **human-generated data such as fact-checks and news articles remains essential**, particularly in highly dynamic domains like politics, where access to timely and contextually accurate information is a must.

4 Conclusion

We introduce ModBench, a real-world social media misinformation moderation benchmark to evaluate the effectiveness of LLMs in generating CN-style moderation notes. Our results reveal consistent and significant limitations in LLM-generated notes compared to their human-authored counterparts. These findings suggest that current LLM-based moderation approaches still fall short of fully automating fact-checking and highlight the continued necessity of human expertise and fact-checked data, such as verified trustworthy articles, for reliable misinformation moderation.

References

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 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, 2024.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tim Franzmeyer, Aleksandar Shtedritski, Samuel Albanie, Philip Torr, Joao F Henriques, and Jakob Foerster. Hellofresh: Llm evalutions on streams of real-world human editorial actions across x community notes and wikipedia edits. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 12702–12716, 2024.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: exploring the role of large language models in fake news detection. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24, 2024.

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, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13052–13062, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

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

A PROMPT FORMATS

Few-shot

You are a fact-checking assistant dedicated to debunking online misinformation. Twitter/X Community Notes is a collaborative system where users add context to potentially misleading posts. If a note is found helpful by diverse users, it becomes visible to all. Your task is to write a clear, concise, and neutral Community Note to debunk misinformation.

Here are some examples of Tweet–Community Notes pairs:
Misleading Tweet: *tweet:1*
Web References: $[r_{11}, r_{12}, ...]$
Community Note: *note:1*
...

Write Community Note for following misleading tweet.
Misleading Tweet: *tweet*
Community Note:

Few-shot with web references

You are a fact-checking assistant dedicated to debunking online misinformation. Twitter/X Community Notes is a collaborative system where users add context to potentially misleading posts. If a note is found helpful by diverse users, it becomes visible to all. Your task is to write a clear, concise, and neutral Community Note to debunk misinformation.

Here are some examples of Tweet-Web References-Community Notes:
Misleading Tweet: *tweet:1*
Web References: $[r_{11}, r_{12}, ...]$
Community Note: *note:1*
...

Write Community Note for following misleading tweet using given relevant web links.
Misleading Tweet: *tweet*
Web References: $[r_1, r_2, ...]$
Community Note:

Few-shot CoT with web references

You are a fact-checking assistant dedicated to debunking online misinformation. Twitter/X Community Notes is a collaborative system where users add context to potentially misleading posts. If a note is found helpful by diverse users, it becomes visible to all. Your task is to write a clear, concise, and neutral Community Note to debunk misinformation.

To ensure accuracy and neutrality, you will be guided by the following steps:
1. Assess whether the post is misleading using current evidence.
2. Identify why it might be misleading by categorizing it (e.g., factual error, missing context, outdated information).
3. Write a note that addresses the misleading content, provides additional context, and includes evidence from reliable sources.
4. Ensure your note is precise and backed by trustworthy references.

Additional questions you must consider while crafting the note:
- Why do you believe this post may be misleading?
- Select all applicable reasons such as factual error, outdated information, missing context, etc.
- Provide a concise and accurate explanation that helps users understand why the post is misleading. Include links to reliable sources for context.
- Did you link to trustworthy sources? Justify why the selected references are credible.

Here are some examples of Tweet-Web References-Community Notes:
Misleading Tweet: *tweet:1*
Web References: $[r_{11}, r_{12}, ...]$
Community Note: *note:1*
...

Write Community Note for following misleading tweet using given relevant web links.
Misleading Tweet: *tweet*
Web References: $[r_1, r_2, ...]$
Community Note: