10^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2024, Philadelphia, USA

Enhancing Crowd-Sourced Fact-Checking Using LLM-Driven Consensus

Keywords: LLM, Community Notes, Crowd Consensus, Misinformation, Jury Learning

Extended Abstract

Crowd-sourced fact-checking has recently been proposed as an effective approach to detect and reduce the spread of misinformation on social media [6, 5, 1, 7]. Twitter's *Community Notes* system is the first large-scale deployment of a crowd-sourced fact-checking system. The system allows users to author *notes* on potentially misleading posts and to rate each other's notes. If a note is rated helpful by enough users with diverse views, it is attached to the post and displayed every time anyone views the post. Twitter deployed the system to all users in December 2023, and so far users have contributed over 620K notes (on more than 377K tweets) and 37M ratings.

For a crowd-sourced fact-checking system like Community Notes to be successful in limiting the spread of misinformation, the fact-checking notes need to be rated helpful soon after the tweet is posted and before it spreads throughout the network. However, initial analysis of the Community Notes data suggests that users tend to write more negative fact-checking notes to posts by counter-partisans [2] and that the time it takes for notes to be deemed helpful is typically not short enough to impact the spread of misinformation significantly [4]. Furthermore, our analysis of the data finds that notes typically fail to be rated helpful quickly because (a) key information is spread across several notes but no single note addresses all inaccuracies in the post, or (b) the notes use language that could be interpreted as biased or argumentative.

In this work, we propose speeding up the time it takes for fact-checking notes to be rated helpful by generating **supernotes** that combine key information from several existing notes using language acceptable to a diverse set of users. A *supernote* is an LLM-generated summary of existing crowd-sourced fact-checks (i.e., notes), which synthesizes useful information and follows key principles of effective fact-checking. Our study is motivated by two key observations: (a) useful information contained in *notes* that are not rated helpful yet are currently underutilized and (b) recent advancements in LLMs can help drive consensus among diversely-opinionated people [3]. We hypothesize that *supernotes* will be rated helpful more widely and more quickly as they are designed to bridge diverse perspectives and explicitly obey key fact-checking principles. This, by extension, should decrease the time it takes for notes to appear publicly and potentially reduce engagement with misinformation.

Generating supernotes. Our pipeline creates an LLM-generated *supernote* using all existing notes and additional information about their ratings. At a high level, we use GPT-4 to generate many candidate supernotes based on all existing notes, and score them based on whether a simulated jury of users would rate them as helpful and whether they follow key principles of effective fact-checking. This pipeline is described further below, and illustrated in Figure 1.

We prompt GPT-4 to generate 100 candidate summaries for each misleading post, using the post's text, all existing community notes, and their current ratings. Then, we evaluate each summary on two key aspects: (a) whether they would be rated helpful by a diverse set of users, and (b) whether they follow first-order principles of effective fact-checking.

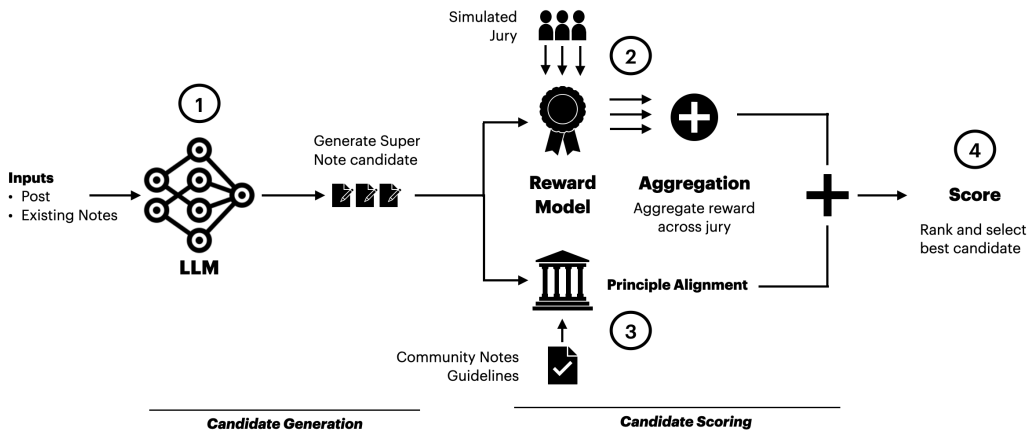

Figure 1: Supernote generation pipeline. (1) We prompt an LLM to generate many candidate summaries using the post text and all existing community notes. (2) We score each candidate summary using a reward model to predict whether a jury of raters would rate the candidate summary as helpful. (3) We filter out candidate summaries that do not follow key principles of effective fact-checking. (4) Finally, we rank and select the summary with the highest score.

Note helpfulness. Recall that for a note to be shown on Twitter it needs to be rated helpful by a diverse set of Community Notes users. We leverage the millions of user ratings in the public Community Notes dataset to build a *reward model* that predicts whether a given user would rate a note as helpful. We train a deep neural network that takes as input a candidate summary embedding (obtained via OpenAI's `ada-002`) and a user embedding (obtained via the Community Notes model) and predicts the user's rating. Our model achieves 69.6% accuracy (AUC: 0.765) on out-of-sample predictions of historical note ratings. Then, we construct a jury of randomly sampled users, predict their individual helpfulness ratings, and aggregate them into a single score using Community Notes' matrix factorization algorithm [8] which scores notes high if they are rated helpful by a diverse set of users.

Principle alignment. In addition to being considered helpful, fact-checking notes must also align with first-order principles of accurate and effective fact-checking. For instance, supernotes must not introduce any information that is not already expressed in existing notes. The Community Notes guidelines provide additional principles that notes should be clear and easy to understand, use unbiased language, be related to claims made in the post, not include opinions or speculations, cite high-quality sources, etc. We prompt GPT-4 to test whether the generated candidate summaries are aligned with these principles and filter out all candidate summaries that are not aligned.

Encouraged by the performance of the individual components of our supernotes generation pipeline, we are currently planning human evaluation experiments. We plan to recruit a diverse group of participants and ask them to compare supernotes against existing notes on the platform. If this evaluation is very promising, we will consider creating a Community Notes account and posting highly-rated supernotes to test their effectiveness in the wild. We hope to highlight several crucial findings that may directly impact and improve the Community Notes program on Twitter.

References

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

[2] Jennifer Allen, Cameron Martel, and David G Rand. 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*, pages 1–19, 2022.

[3] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

[4] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. The roll-out of community notes did not reduce engagement with misinformation on twitter. *arXiv preprint arXiv:2307.07960*, 2023.

[5] Hyunuk Kim and Dylan Walker. Leveraging volunteer fact checking to identify misinformation about covid-19 in social media. *HKS Misinformation Review*, 2020.

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

[7] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. Searching for or reviewing evidence improves crowdworkers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, 2(2):26339137231173407, 2023.

[8] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*, 2022.