COMMUNITYNOTES: A Dataset for Exploring the Helpfulness of Fact-Checking Explanations

Anonymous ACL submission

Abstract

Fact-checking on major platforms, such as X, Meta, and TikTok, is shifting from expert-driven verification to a community-based setup, where users contribute explanatory notes to clarify why a post might be misleading. An important challenge here is determining whether an explanation is helpful for understanding real-world claims and the reasons why, which remains largely underexplored in prior research. In practice, most community notes remain unpublished due to slow community annotation, and the reasons for helpfulness lack clear definitions. To bridge these gaps, we introduce the task of predicting both the helpfulness of explanatory notes and the reason for this. We present COMMUNITYNOTES, a large-scale multilingual dataset of 105k posts with user-provided notes and helpfulness labels. We further propose a framework that automatically generates and improves reason definitions via automatic prompt optimization, and integrate them into prediction. Our experiments show that the optimized definitions can improve both helpfulness and reason prediction. Finally, we show that the helpfulness information are beneficial for existing fact-checking systems.[1]

1 Introduction

Beyond simple verification, fact-checking systems should provide explanations that clarify why a claim is misleading (Warren et al., 2025). Explanations not only strengthen user trust, but also mitigate pitfalls of fact-checking, e.g., backfire, where the false belief is reinforced rather than hindered (Lewandowsky et al., 2012; Guo et al., 2022; Xing et al., 2025). Typically, fact-checking explanations are produced by expert organizations. Yet the ecosystem is currently shifting, and platforms are increasingly experimenting with community-based fact-checking, where contributors collaboratively provide explanatory notes on claims (Augenstein et al., 2025). This crowdsourcing process distributes the labor of verification and offers diverse perspectives from the community. A representative example is *Community Notes* on X (formerly Twitter), which allows users to attach notes explaining potentially misleading posts (Figure 1).[2]

Figure 1: An example of X Community Notes. The user-generated note appears as "*Readers added context*."

Critically, for a note to become visible or formally published, it must receive sufficient amount of ratings from a diverse set of users (De et al., 2025). A key challenge in that respect is to decide whether an explanation is helpful for understanding real-world claims and why, which is underexplored in previous research on fact-checking explanations (Atanasova et al., 2020; Guo et al., 2022; Russo et al., 2023; Eldifrawi et al., 2024). This

[1] Code and data are available at: ANONYMIZED.

[2] https://communitynotes.x.com/guide/

process is also slow on *Community Notes*: notes can take from hours to days to appear, and the vast majority (over 90%) never become visible, despite the invested contributors efforts. Moreover, while the platform provides reasons for judging notes as helpful or unhelpful, these reasons lack clear definitions, leaving both raters and contributors uncertain about the evaluation criteria (De et al., 2025; Augenstein et al., 2025). This opacity hinders the efficiency, scalability, and explainability of crowd-sourced fact-checking.

Motivated by the above, here we explore the idea of simplifying and making more explicit the helpfulness annotation process. Our contributions are as follows:

- We introduce the task of predicting the helpfulness and its reasons given a note and the original X post.
- We present COMMUNITYNOTES, a large-scale explanation helpfulness dataset containing over 105K potentially misleading posts and corresponding user-provided notes.
- We propose a framework that automatically generates and optimizes reason label definitions via prompt optimization; integrating them to helpfulness/reason predictors to improve their performance.
- We apply helpfulness prediction to the evidence sufficiency task (Atanasova et al., 2022) and further incorporate it as auxiliary signals into automated fact-checking on CLIMATE-FEVER (Diggelmann et al., 2020), showing its potential to enhance existing fact-checking systems.

2 Related Work

Community-Based Fact-Checking The proliferation of misinformation has become a growing societal concern, posing significant risks to areas such as public health, science, and politics (Borenstein et al., 2025). To address this challenge, fact-checking has emerged as a central strategy in mitigating the spread of false or misleading information (Guo et al., 2022).

Traditionally, fact-checking has been conducted by professional third-party organizations that evaluate the veracity of claims through expert analysis. However, recent years have witnessed a paradigm shift toward more decentralized, community-based approaches (Renault et al., 2024; Augenstein et al., 2025). Major social media platforms have increasingly embraced this model. X's *Community Notes* has become the most prominent example, enabling users to collaboratively annotate and assess the accuracy of online content. Following this trend, Meta recently terminated its partnerships with professional fact-checkers in favor of a similar community-driven system (Catalanello and Sanders, 2025), while TikTok has also introduced a comparable feature (TikTok, 2025).

Empirical evidence suggests that community-based fact-checking can be highly effective in countering misinformation. Studies have shown that the presence of a Community Note can reduce the rate of retweets by nearly half (Renault et al., 2024), while displaying such notes on misleading posts has been found to decrease their spread by an average of 61% (Chuai et al., 2024). Moreover, authors of posts flagged by community notes are up to 80% more likely to delete their content, and such posts generally exhibit markedly lower virality (Drolsbach and Pröllochs, 2023; Renault et al., 2024). However, *Community Notes* often requires a long time to gather sufficient contributor ratings to determine helpfulness, and the absence of reason definitions further undermines the trustworthiness of the mechanism.

Label Definitions Recent work has shown that providing models with explicit definitions of labels can lead to substantial improvements in classification. Khatuya et al. (2025) used predefined label descriptions and trained the model to generate these descriptions based on the input text, achieving sizable improvements in multi-label classification. Similarly, Gao et al. (2023) demonstrated that curating a small dictionary of terms for label description can improve zero-shot accuracy by 17-19% absolute, while also increasing the robustness to prompts. Peskine et al. (2023) showed improvements when using label definitions with GPT-3 zero-shot classification on the challenging task of fine-grained conspiracy theory detection.

Automatic Prompt Optimization Prompt engineering has become essential to many Natural Language Processing (NLP) tasks; yet, manual trial-and-error approaches remain a barrier to effective use. Automatic Prompt Optimization (APO) methods address this challenge by systematically searching for high-performing prompts (Ramnath et al., 2025). These approaches explore the

large instruction space using diverse strategies: PROMPTBREEDER evolves task-specific prompts across generations using optimized mutation strategies (Fernando et al., 2024); PROMPTAGENT applies Monte Carlo Tree Search (MCTS) with feedback reflection to guide exploration (Wang et al., 2024); and SCULPT represents prompts as trees, enabling targeted refinements while preserving overall structure—especially useful for long or complex prompts (Kumar et al., 2025).

3 Predicting Whether an Explanation is Helpful and Why

3.1 Task Definition

Given a post P that contains potentially misleading information and a note N that provides an explanation about that post, the task is to predict two types of labels: a **helpfulness label** $L_{helpful}$, indicating whether the note is helpful in explaining the misleading nature of the post, and a **reason label** L_{reason}, specifying why the note is (un)helpful. The helpfulness prediction is formulated as a binary classification task with the label space $\{Helpful, Unhelpful\}$. The reason prediction is a multi-label classification task, where the predefined label set includes eight reasons for helpful notes and ten reasons for unhelpful ones (See Appendix A.2 for detailed reason labels).

3.2 Data Collection and Pre-processing

We collected all our data from official *Community Notes* website between January 2021 and December 2024.[3] We construct our dataset by combining all publicly available data releases, which include the Notes, Ratings, Note Status History, and User Enrollment tables. Since the ratings data are provided in multiple shards due to their large volume, we first merge all shards and join them with the other components to obtain complete note–rating pairs. We then apply the official note-ranking algorithm to compute the aggregated helpfulness and reason labels for each note.[4] For helpfulness prediction, we remove entries labeled as NEED_MORE_RATINGS, leaving the helpfulness label space binary—CURRENTLY_RATED_HELPFUL or CURRENTLY_RATED_NOT_HELPFUL. The reason annotation is multi-label, covering 18 predefined categories. Finally, we retrieve tweet content and metadata via the X API and web crawling (Borenstein et al., 2025), linking each post with its corresponding notes to form the final dataset used in our experiments.

Language	Train	Dev	Test	Total
English	41,102	5,872	11,744	58,718
Other languages	32,520	4,646	9,292	46,458
Total	73,622	10,518	21,036	105,176

Table 1: Data split statistics for English and other language subsets.

Table 1 provides the data distribution. Our final dataset contains 105,176 posts, including 58,718 in English and 46,458 in other languages. We perform stratified partitioning with a 7:1:2 ratio for training, development, and testing sets, maintaining separate splits for English and other language data. Henceforth we will refer to this dataset as COMMUNITYNOTES.

3.3 Dataset Statistics

Text Type	Mean	Median	Min	Max
Post	57.44	42	0	5,072
Note	87.05	70	1	1,402

Table 2: Token length statistics for claims and notes. Texts are tokenized by OpenAI's tiktoken v0.11.0.

We inspected multiple aspects of our entire COMMUNITYNOTES dataset. Table 2 shows the token length distribution for posts and notes correspondingly. The median token number are 42 for posts and 70 for notes, which suggests most posts and notes are short text snippets.

Notes per Post	Count (%)
1	84.74%
2	11.84%
3	2.43%
4	0.65%
5	0.18%
6	0.08%
7	0.03%
8	0.01%
$\geq$9	$<$0.02%

Table 3: Distribution of the number of notes per post. The majority of posts have only one associated note.

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

[4] https://communitynotes.x.com/guide/en/under-the-hood/ranking-notes

Post Type	Count	Percentage
All helpful notes	67,523	64.20%
All unhelpful notes	33,061	31.43%
Mixed helpful / unhelpful notes	4,592	4.37%

Table 4: Distribution of posts by helpfulness composition of their associated notes.

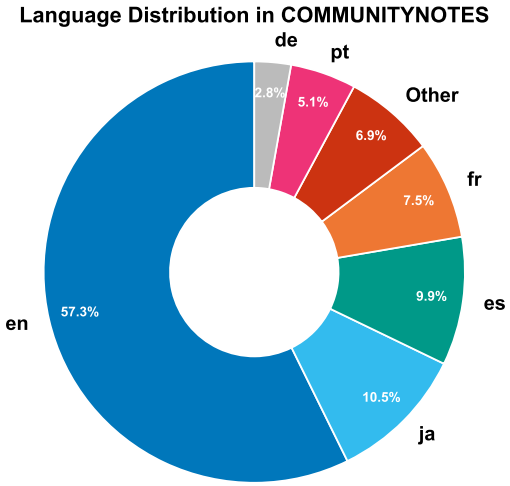

Figure 2: The languages in the COMMUNITYNOTES: fr-French, es-Spanish, ja-Japanese, pt-Portuguese, de-German, Other: note languages that appear less than 1,000 times.

In Table 3, we observe that 84.74% of posts have only one note, 11.84% have two notes, and the remaining posts contain three or more notes. Table 4 shows that 64.2% of posts have notes that are all rated as helpful, 31.4% have only unhelpful notes, and the remaining 4.4% include notes with mixed ratings. There are eight predefined reasons for a note being rated as helpful and ten reasons for being rated as unhelpful. The most common helpful reasons are `helpfulAddressClaim`, `helpfulImportantContext`, and `helpfulClear`, whereas `nothelpfulArgumentativeOrBiased` and `nothelpfulMissingKeyPoints` are the most frequent unhelpful reasons (See detailed reason distribution in Appendix Figure 4).

Figure 2 illustrates the language distribution of COMMUNITYNOTES. English is the dominant language, accounting for 57.3% of all notes, followed by Japanese (10.5%) and Spanish (9.9%). This indicates that our final dataset effectively captures the multilingual nature of crowd-sourced explanations.

Beyond analyzing note statistics, we also examine the popularity of the posts targeted by these notes. We use the number of replies, likes, and retweets as metrics (See detailed distributions in Appendix Figure 5). These posts receive a median of approximately 3k likes, 300 replies, and 700 retweets. This indicates that the posts attracting community notes are highly popular. Given that they are flagged as potentially misleading, this underscores the urgent need to provide helpful explanations for their claims.

3.4 Benchmarking

Experimental Setup To establish a baseline for understanding the model performance in predicting the helpfulness of the notes, we conducted a comprehensive set of experiments using both Small Language Models (SLMs) (BERT-style encoder-only architectures in our paper) and decoder-only Large Language Models (LLMs). For SLMs, we formulate the task as multi-task classification, where the model is jointly trained to predict (1) a binary helpfulness label and (2) a multi-label reason classification across 18 predefined categories. The model input is constructed as a concatenated sequence in the format "[CLS] Claim: {claim} [SEP] Note: {note_text}" and subsequently passed through two separate classification heads. For LLMs, we fine-tune models in the same multi-task way (See detailed prompt in Appendix Appendix A.2).

We partitioned our experiments by model and language setting. For English, we experimented with SLMs including BERT, ModernBERT, RoBERTa and DeBERTa (both base and large). We also evaluated the performance of LLMs including Llama3.1-8B and Mistral-7B, under both zero-shot and LoRA fine-tuning conditions (Hu et al., 2022). For the multilingual part, we used models specifically trained on multilingual data, including BERT-base-multilingual (mBERT) and XLM-RoBERTa (base and large).

Results and Analysis The models generally perform well on the helpfulness prediction task, with most of them achieving F1 scores beyond 0.88. This suggests that distinguishing between helpful and non-helpful notes is a relatively easy task. In contrast, the reasons prediction task is much more challenging. This is showed by the much lower F1 scores, particularly the F1 scores which consistently fall below 0.70.

Comparing SLMs to LLMs: The table shows that LLMs like Llama3.1-8B and Mistral-7B outperform smaller SLMs on Helpfulness predic-

Language	Model	Helpfulness F1	Reason P	Reasons R	Reason F1
English	BERT-base	0.866	0.549	0.742	0.631
	BERT-large	0.874	0.553	0.758	0.640
	RoBERTa-base	0.876	0.546	0.783	0.643
	RoBERTa-large	0.890	0.581	0.772	0.663
	ModernBERT-base	0.871	0.545	0.755	0.633
	ModernBERT-large	0.888	0.576	0.752	0.652
	DeBERTa-base	0.886	0.541	**0.810**	0.650
	DeBERTa-large	0.896	0.589	0.759	**0.665**
	Llama3.1-8B-ins	0.919	**0.640**	0.640	0.640
	Mistral-7B-ins	**0.920**	0.620	0.620	0.620
Other	mBERT	0.873	0.562	0.799	0.659
	XLM-RoBERTa-base	0.882	0.584	0.798	0.674
	XLM-RoBERTa-large	**0.894**	**0.603**	**0.815**	**0.693**

Table 5: Benchmark results for helpfulness and reasons prediction. *Helpfulness* means the binary classification of notes helpfulness. *Reason* means multi-label classification of helpfulness reasons.

tion. The Mistral-7B model achieves the highest F1 Score (0.920), demonstrating the superior capabilities of larger, more advanced models. However, the LLMs perform worse than SLMs in reason prediction, with their F1 scores being comparable to the best-performing SLMs like DeBERTa-large. Within the SLM families (BERT, RoBERTa, DeBERTa), larger models consistently outperform their base-sized counterparts. For example, RoBERTa-large achieves better scores than RoBERTa-base across all metrics. This highlights the importance of model capacity for capturing the nuances required for this complex task. The table also shows that DeBERTa models generally perform at or near the top among the SLMs, particularly for the reasons task with the highest F1 Score of 0.665 among all SLMs.

The multilingual models (mBert and XLM-RoBERTa) show strong performance, especially on the helpfulness prediction task. XLM-RoBERTa-large achieves the highest Helpfulness F1 of 0.894. This indicates that these models are effective at generalizing across different languages.

4 Enhancing Helpfulness Prediction with Reason Definitions

Preliminary analysis shows that predicting why a particular explanation is (un)helpful is often subjective across instances, largely due to the lack of unified or official definitions. This subjectivity poses a barrier to model performance and limits the explainability of community-based fact-checking platforms. Prior work has demonstrated that providing clear label definitions can significantly improves the classification performance (Gao et al., 2023; Peskine et al., 2023; Khatuya et al., 2025). Given the large number of notes with annotated reason available, we aim to automatically generate and optimize reason label definitions based on data. To this end, we introduce a pipeline that leverages Automatic Prompt Optimization (APO) to generate and improve helpfulness reason definitions, which are then integrated with notes for reason prediction. Since the majority of notes and definitions are written in English, we focus on **English** data for the remainder of our experiments. Figure 3 illustrates the overall pipeline.

4.1 Automatically Optimizing Reason Definition

We generate and optimize reason definitions following the first two steps in Figure 3: (1) *Seed reason definition generation* and (2) *Reason definition optimization*. In the first step, we follow Peskine et al. (2023) and randomly sample 40 instances for each category. We then prompt the GPT-4o model to generate candidate definitions (see Appendix A.2 in Appendix for detailed prompts).[5]

In the second step, we employ PROMPTAGENT (Wang et al., 2024), a Monte Carlo Tree Search (MCTS)-based automatic prompt optimiza-

[5]Since a single sample may have multiple reason labels, we perform an additional check to ensure that each sample appears only within the same category to encourage diversity.

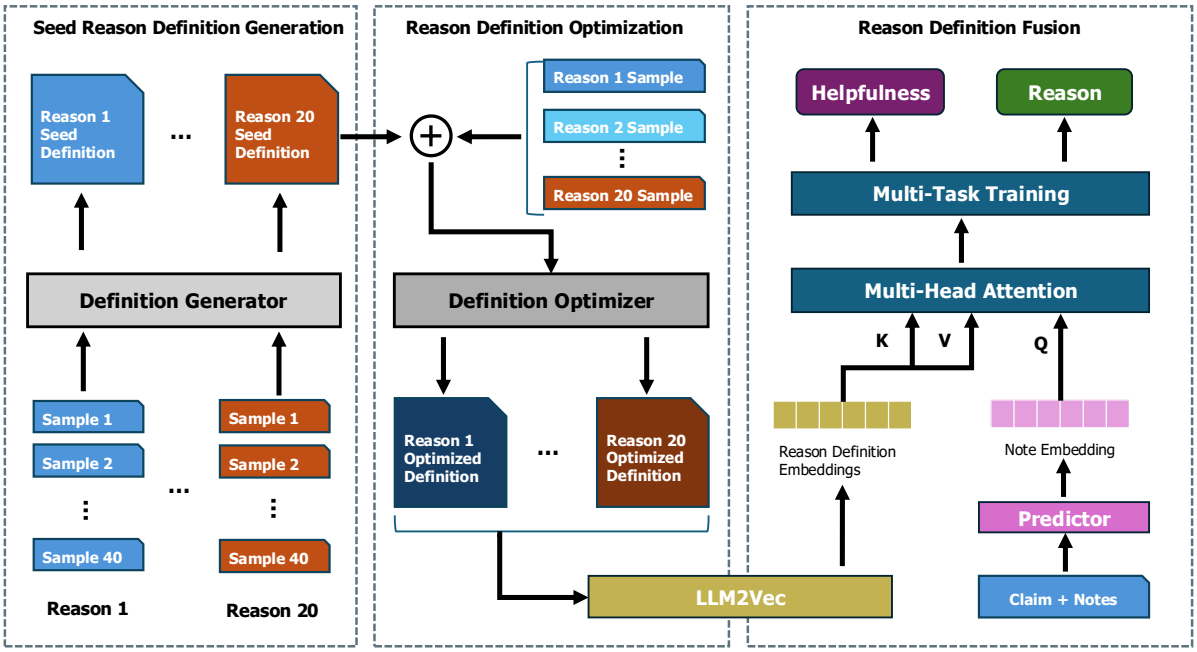

Figure 3: Our framework for automatic reason definition generation and optimization.

tion framework, to refine the initial definitions. PROMPTAGENT adopts an iterative refinement process that emulates how human experts strategically craft superior prompts, implemented through an MCTS-based planning mechanism. Each refined prompt corresponds to a new state in the search space, obtained by the refiner agent model through feedback collected from a feedback agent model. PROMPTAGENT relies on the MCTS state–action value function to estimate the potential reward of each state. MCTS iteratively performs four key operations—selection, expansion, simulation, and backpropagation—until a predefined number of iterations is reached. The highest-reward trajectory is then selected to produce the final refined prompt. We initialize the process with the seed definitions generated in step (1) and run PROMPTAGENT using a definition-refinement feedback loop.

To evaluate the impact of incorporating generated definitions, we perform zero-shot prompting with Llama3.1-8B-Instruct[6] and Mistral-7B-Instruct[7] for both helpfulness and reason prediction. Table 6 reports results for the vanilla setting, as well as when incorporating seed and optimized definitions. We observe substantial F1 improvements in both helpfulness and reason prediction when

Model	Helpfulness F1	Reason F1
Llama3.1-8B-ins	0.636	0.126
Llama3.1-8B-ins-seed	0.783	0.304
Llama3.1-8B-ins-opt	**0.785**	**0.344**
Mistral-7B-ins	0.761	0.135
Mistral-7B-ins-seed	0.776	0.257
Mistral-7B-ins-opt	**0.786**	**0.324**

Table 6: Zero-shot prediction performance for LLMs w/o adding reason definitions on our COMMUNITYNOTES test set. *No suffix* means the vanilla model prediction performance. *seed* means definitions are generated by GPT-4o on randomly sampled cases. *opt* means that the definitions are further optimized by PROMPTAGENT.

adding seed definitions (on average around 15%), with further gains after definition optimization. For Llama3.1-8B-Instruct, both seed and optimized definitions yield large improvements. Although Mistral-7B-Instruct starts stronger than Llama3.1-8B-Instruct in the vanilla setting, it still benefits notably from incorporating definitions, particularly in reason-level F1. These results suggest that incorporating reason definition information substantially enhances model performance—not only in predicting whether a given explanation note is helpful, but also in identifying the underlying reason with respect to the claim (See seed definition in Appendix A.2 and optimized reason definition in

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Appendix Table 15).

4.2 Training with Reason Definition Fusion

After obtaining the optimized helpfulness reason definitions, we proceed to the third step in Figure 3, where we explore how to incorporate this information into the predictor's training process. The core idea is that, for a given note associated with a claim, the predictor should attend more strongly to the relevant helpfulness reasons. Since SLM performs better in reason prediction, we use SLM as our predictors in the experiments.

We first obtain note embeddings E_{note} from the original predictors and reason embeddings E_{reason} from the optimized reason definitions using LLM2VEC (BehnamGhader et al., 2024), an unsupervised LLM-based embedding model.[8] We then apply a Multi-Head Attention (MHA) mechanism, using E_{note} as the query (Q) and E_{reason} as the key (K) and value (V), to fuse the reason information into the note representation. Finally, we adopt a multi-task learning objective that jointly optimizes helpfulness prediction and multi-label reason classification.

Table 7 shows the impact of integrating the optimized reason definitions into the base models using the MHA module. Across all predictors, incorporating reason definitions via MHA consistently improves both helpfulness and reason F1 scores. For instance, RoBERTa-large-MHA achieves a Helpfulness F1 of 0.899 and a Reason F1 of 0.669, outperforming its vanilla counterpart (0.890 and 0.663, respectively). Similar trends hold for ModernBERT and DeBERTa, with DeBERTa-large-MHA achieving the highest Reason F1 score of 0.677. These results demonstrate that attention-based fusion of reason definitions effectively enhances SLMs, improving their ability to predict both whether a note is helpful and the underlying reason.

5 Generalization of the Explanation Helpfulness

In this section, we assess the generalization and the utility of our helpfulness predictors beyond community notes. A key challenge is that most existing datasets lack explicit annotations for explanation of helpfulness. Therefore, we explore two key scenarios in automated fact-checking. First, we evaluate whether helpfulness prediction can generalize to evidence sufficiency prediction, which is an important subtask in fact-checking. Second, we explore the effect of incorporating evidence helpfulness in automated fact-checking.

[8] Specifically, we use `LLM2Vec-Meta-Llama-31-8B-Instruct-mntp-supervised`.

Model	Helpfulness F1	Reason F1
RoBERTa-base	0.876	0.643
RoBERTa-base-MHA	**0.887**	**0.652**
RoBERTa-large	0.890	0.663
RoBERTa-large-MHA	**0.899**	**0.672**
ModernBERT-base	0.871	0.633
ModernBERT-base-MHA	**0.874**	**0.644**
ModernBERT-large	0.888	0.652
ModernBERT-large-MHA	**0.898**	**0.662**
DeBERTa-base	0.886	0.650
DeBERTa-base-MHA	**0.887**	**0.668**
DeBERTa-large	0.896	0.665
DeBERTa-large-MHA	0.896	**0.677**

Table 7: Performance comparison of base models versus fusing optimized reason definition with the Multi-Head Attention module.

5.1 Generalization to Evidence Sufficiency

Evidence sufficiency is a task that decides whether the information in the evidence is sufficient for predicting the veracity of a claim. Automated fact-checking models can make veracity predictions only when there is enough evidence (Atanasova et al., 2022). In this scenario, we would like to explore whether our helpfulness predictors can predict the sufficiency of an evidence (since a helpful evidence must be "sufficient"). We adopted *Sufficient Fact* (Atanasova et al., 2022), a diagnostic fact-checking dataset to detect when evidence with omitted information is (in)sufficient. To adapt our prediction, we treat sufficient evidence (*EI-ENOUGH INFO*) as helpful and insufficient evidence (*NEI-NOT ENOUGH INFO*) as unhelpful.

Table 8 shows the model performance on the evidence sufficiency task, focusing on the *NEI* label. As expected, fine-tuned (FT) models outperform the original notes predictors, with RoBERTa-large-FT achieving the best *NEI* F1 of 0.908. Nevertheless, the notes models show competitive performance: for example, ModernBERT-large-notes achieves a *NEI* F1 of 0.808, only about 10% lower than its FT counterpart. Across all predictors, the top predicted reason is *notHelpfulMissingKeyPoints*, which closely corresponds to the *NEI* label, as both signify the absence of critical information.

Predictors	NEI P	NEI R	NEI F1
RoBERTa-base-FT	0.869	0.920	0.894
RoBERTa-base-notes	0.742	0.521	0.612
RoBERTa-large-FT	0.882	0.935	0.908
RoBERTa-large-notes	0.731	0.676	0.702
ModernBERT-base-FT	0.845	0.843	0.844
ModernBERT-base-notes	0.770	0.779	0.774
ModernBERT-large-FT	0.902	0.906	0.904
ModernBERT-large-notes	0.808	0.807	0.808
DeBERTa-base-FT	0.912	0.855	0.882
DeBERTa-base-notes	0.746	0.553	0.635
DeBERTa-large-FT	0.919	0.890	0.905
DeBERTa-large-notes	0.736	0.774	0.755

Table 8: Predictor generalization performance on evidence sufficiency task. *notes* stands for models for original communitynotes predictors. *FT* means **F**ine**T**uned models on evidence sufficiency task. **NEI** means **N**ot **E**nough **I**nformation label.

This indicates that helpfulness prediction carries transferable signal for evidence sufficiency. Still, helpfulness and sufficiency are not the same: even sufficient evidence could be predicted as unhelpful for other reasons such as *notHelpfulArgumentativeOrBiased*, which may account for the observed performance gap.

5.2 Incorporating Evidence Helpfulness in Automated Fact Checking

In this part, we further investigate whether predicted evidence helpfulness can enhance automated fact-checking. Helpfulness scores provide auxiliary signals for assessing evidence quality, with the potential to help models prioritize informative evidence during prediction. We adopted `CLIMATE-FEVER` (Diggelmann et al., 2020), a real-world dataset for verifying climate change–related claims. Each claim is associated with multiple evidence sentences, to which we apply our predictors to assign helpfulness scores and reasons. To perform fact-checking, we use Llama3.1-8B-Instruct to perform fact-checking (See detailed prompt in Appendix Table 16) prompting the model with both the claim and its associated evidence.

We compare the accuracy with and without the helpfulness information of evidence and report results in Table 9. We can see that, in general, incorporating helpfulness improves fact-checking performance with ModernBERT-base achieved an accuracy of 0.537. ModernBERT and RoBERTa-base improve fact-checking performance by around 2%. These findings indicate that providing auxiliary evidence helpfulness and reasons can benefit automated fact-checking. While ModernBERT demonstrates major gains, DeBERTa and RoBERTa yield smaller, non-significant improvements. Interestingly, this pattern mirrors their performance drop in cross-domain evaluation of evidence sufficiency in Section 5.1 (despite having the best in-domain results in Table 5). One explanation could be that these models have overfitted for the *Community Notes* domain and as such are less effective in other domains.

Predictors	Acc	+Helpfulness Acc
RoBERTa-base	0.513	**0.521**
RoBERTa-large	0.519	0.521
ModernBERT-base	0.515	**0.537**
ModernBERT-large	0.517	**0.535**
DeBERTa-base	0.518	0.519
DeBERTa-large	0.519	0.519

Table 9: Automated fact-checking performance of Llama3.1-8B-instruct on the `CLIMATE-FEVER` dataset. *Acc* stands for Accuracy. *Acc +Helpfulness* means accuracy incorporating helpfulness information. Numbers in bold represents the gap between *+Helpfulness* and basic accuracy is statistically significant (p-value <0.05).

6 Conclusion

In this paper, we introduced the novel task of predicting the helpfulness of explanations and their underlying reasons for a given claim. To support this task, we constructed `COMMUNITYNOTES`, a large-scale post–note helpfulness dataset based on X's *Community Notes*, and established benchmarks with pretrained language models. While existing models perform reasonably well on overall helpfulness, they struggle to identify underlying reasons. To address this, we proposed a framework that automatically generates and optimizes reason definitions, which, when reintegrated, greatly improves reason prediction. We further demonstrated that helpfulness prediction can help predict evidence sufficiency and enhance fact-checking systems.

We hope that our dataset and framework can serve as a foundation for building stronger helpfulness predictors capable of assigning prior labels to real-time community notes, thereby streamlining the community-based fact-checking process. Additionally, the automatically generated reason definitions may enhance the explainability and trustworthiness of fact-checking models and could be extended to improve other classification tasks beyond this domain.

Limitations

While our study provides new insights into the helpfulness of crowd-sourced fact-checking explanations, several limitations remain.

First, the inherently subjective nature of community-based annotations introduces potential noise and inconsistency in the helpfulness and reason labels. Despite our efforts to aggregate ratings through official mechanisms, differences in annotator interpretation and cultural context may still influence the results.

Second, our automated definition generation and optimization pipeline has thus far been evaluated primarily on English. Although our dataset includes multilingual content, we did not systematically validate the quality or transferability of the generated definitions across languages. Future work should explore multilingual adaptation and cross-lingual alignment of reason definitions.

Third, while our experiments show improved model performance when integrating optimized reason definitions, the evaluation is limited to classification metrics. Further human-centered evaluation is necessary to assess whether these improvements translate to greater interpretability and practical usefulness in real-world fact-checking workflows.

Finally, our study focuses on publicly available *Community Notes* data, which may reflect platform-specific norms and biases. For a broader generalization to other fact-checking ecosystems (e.g., Meta's or TikTok's community moderation systems), we need further investigation.

Ethics and Broader Impact

Our work aims to improve the transparency, explainability, and scalability of community-based fact-checking systems. By modeling and predicting the helpfulness of user-generated explanations, we hope to assist both contributors and platforms in identifying notes that are more informative, balanced, and useful for the public discourse. However, several ethical considerations deserve discussion.

Bias and Fairness. Community-based ratings may reflect social, cultural, or political biases that influence which notes are deemed "helpful." Our models inevitably inherit such biases from the underlying data. We emphasize that model predictions should not be used to automatically moderate or suppress user content. Instead, they should be interpreted as supporting signals to guide human review.

Responsible Use. While our framework can aid in prioritizing or surfacing potentially helpful notes, it is not intended to replace human judgment. Automated systems that assign helpfulness or credibility labels at scale could have downstream impacts on content visibility, user reputation, and the dynamics of online discourse. We encourage future work to incorporate fairness-aware and interpretability-centered mechanisms before deployment.

Positive Societal Impact. Despite these caveats, we believe our dataset and models can contribute positively to combating misinformation by improving the quality, trustworthiness, and accessibility of community fact-checking. By making the definitions of helpfulness explicit and machine-interpretable, we hope to support more transparent, accountable, and participatory approaches to information verification.

Data Collection and Licenses All data used in this work originate from the publicly available *X Community Notes* dataset,[9] which is released by the platform under its open data policy. The dataset consists of user-generated notes, their associated posts, and corresponding helpfulness and reason labels. No personally identifiable information (PII), user handles, or private metadata are included in our version of the dataset. We processed and released only anonymized text and aggregated statistics in compliance with the platform's terms of service. The dataset is multilingual, with English, Japanese, and Spanish as the most frequent languages. All examples are drawn from publicly visible content. Any future release of derived data or annotations will follow ACL's data ethics guidelines and include proper documentation describing preprocessing steps, filtering criteria, and licensing terms.

Security Implication Our study does not introduce direct security risks, but the use of models trained on user-generated content entails potential vulnerabilities. Automated systems that evaluate or rank explanatory notes could be misused for large-scale manipulation—for example, to promote particular narratives or suppress dissenting perspectives. To mitigate such risks, we release only aggregated and anonymized data, and we emphasize

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

that our models are intended for research and analysis, not for automated moderation or deployment without human oversight. Future work should include safeguards against adversarial attacks, bias exploitation, and model inversion risks, particularly when integrating such predictors into live social platforms.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Isabelle Augenstein, Michiel Bakker, Tanmoy Chakraborty, David Corney, Emilio Ferrara, Iryna Gurevych, Scott Hale, Eduard Hovy, Heng Ji, Irene Larraz, Filippo Menczer, Preslav Nakov, Paolo Papotti, Dhruv Sahnan, Greta Warren, and Giovanni Zagni. 2025. Community moderation and the new epistemology of fact checking on social media. *Preprint*, arXiv:2505.20067.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *Proceedings of the First Conference on Language Modeling*.

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

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. 2024. Community Notes Reduce the Spread of Misleading Posts on X.

Soham De, Michiel A. Bakker, Jay Baxter, and Martin Saveski. 2025. Supernotes: Driving consensus in crowd-sourced fact-checking. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3751–3761, New York, NY, USA. Association for Computing Machinery.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

Chiara Patricia Drolsbach and Nicolas Pröllochs. 2023. Diffusion of community fact-checked misinformation on twitter. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6679–6692, Bangkok, Thailand. Association for Computational Linguistics.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13823–13844, Singapore. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*.

Subhendu Khatuya, Shashwat Naidu, Saptarshi Ghosh, Pawan Goyal, and Niloy Ganguly. 2025. Label-semantics aware generative approach for domain-agnostic multilabel classification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22286–22298, Vienna, Austria. Association for Computational Linguistics.

Shanu Kumar, Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, Parag Agrawal, and Manish Gupta. 2025. SCULPT: Systematic tuning of long prompts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14996–15029, Vienna, Austria. Association for Computational Linguistics.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063,

Singapore. Association for Computational Linguistics.

Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, and 2 others. 2025. A systematic survey of automatic prompt optimization techniques. *Preprint*, arXiv:2502.16923.

Thomas Renault, David Restrepo-Amariles, and Aurore Troussel. 2024. Collaboratively Adding Context to Social Media Posts Reduces the Sharing of False News. HEC Research Papers Series 1519, HEC Paris.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

TikTok. 2025. Testing a New Feature to Enhance Content on TikTok.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. 2024. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking.

Rui Xing, Timothy Baldwin, and Jey Han Lau. 2025. Evaluating evidence attribution in generated fact checking explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5475–5496, Albuquerque, New Mexico. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

We present more data statistics about COMMUNITYNOTES in Figure 4 and Figure 5.

A.2 Prompt details

In this section, we provided detailed prompt used in COMMUNITYNOTES experiments.

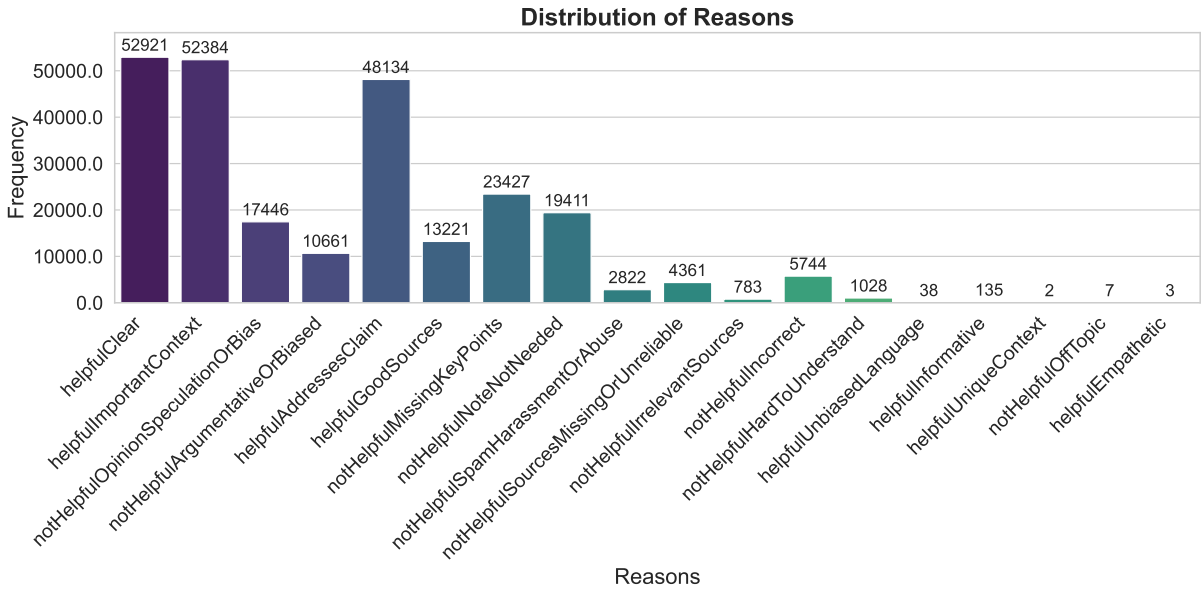

Figure 4: A historgram of the *Note Reason* label frequencies.

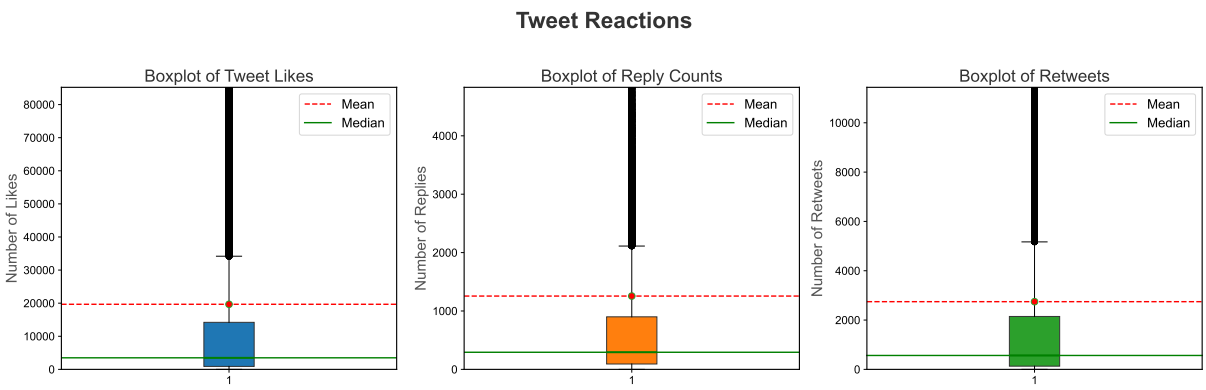

Figure 5: Boxplot of tweet reactions in COMMUNITYNOTES.

Initial reason generation
You will be given a set of samples, each sample contains CLAIM, their corresponding NOTE to explain the CLAIM. All samples provided are ${helpful_label} in explaining the CLAIM and associated with the same REASON. Your task is to conclude the definition of the REASON. Here are samples: ${samples} The REASON for above samples being ${helpful_label} is ${reason_label}. After checking these samples, the definition of this REASON is:

Table 10: Prompt for initial reason generation

Original Prompt
Given a potentially misleading CLAIM and an associated NOTE, your task is to determine whether the NOTE is helpful in clarifying the CLAIM and identify two reasons from the predefined reason set explaining why it is helpful or not helpful. The predefined reason set is:'helpfulAddressesClaim', 'helpfulClear', 'helpfulEmpathetic', 'helpfulGoodSources', 'helpfulImportantContext', 'helpfulInformative', 'helpfulUnbiasedLanguage', 'helpfulUniqueContext', 'notHelpfulArgumentativeOrBiased', 'notHelpfulHardToUnderstand','notHelpfulIncorrect', 'notHelpfulIrrelevantSources', 'notHelpfulMissingKeyPoints', 'notHelpfulNoteNotNeeded', 'notHelpfulOffTopic', 'notHelpfulOpinionSpeculation', 'notHelpfulOpinionSpeculationOrBias', 'notHelpfulOther', 'notHelpfulSourcesMissingOrUnreliable', 'notHelpfulSpamHarassmentOrAbuse' Output in the following JSON format only, no extra output: {"helpfulness": helpful or non_helpful,"reasons":"reason1;reason2"} CLAIM: ${claim} NOTE: ${note} Answer:

Table 11: Original English prompt for note helpfulness and reason prediction.

Prompt with seed defininion
Given a potentially misleading CLAIM and an associated NOTE, your task is to determine whether the NOTE is helpful in clarifying the CLAIM and identify two reasons from the predefined reason set explaining why it is helpful or not helpful. Here are reasons and their definitions: ${reason definitions} Output in the following JSON format only, no extra output: {"helpfulness": helpful or non_helpful,"reasons":"reason1;reason2"} CLAIM: ${claim} NOTE: ${note} Answer:

Table 12: English prompt with seed definition for note helpfulness and reason prediction.

Optimized Prompt
For each claim-note pair below, select exactly two reasons (from the provided list) that most accurately explain whether and why the note is helpful or not helpful for understanding or resolving the claim.

Continued on next page

Optimized Prompt

High-Level Criteria

1. Prioritizing Central, Accurate Clarification

- Assign a "helpful" reason only if the note is factually accurate, well-supported, and directly clarifies, corrects, or provides essential context for the main assertion, a key factual sub-claim, or the foundational evidence of the claim. Peripheral or partially related details are insufficient.
- If the note's chief function is to correct a specific fact, figure, identity, date, source, or a significant misattribution or misconception central to the claim, this counts as clarification or correction (use "helpfulClear" and/or "helpfulAddressesClaim").
- Never assign any "helpful" reason if the note contains substantive factual errors, misrepresentations, speculation, or bias regarding any key claim point or related evidence—regardless of any attempt to clarify.

2. "Helpful" Reason Prioritization

- helpfulClear and helpfulAddressesClaim take precedence over other "helpful" reasons when a note offers a direct factual correction or explicit clarification for the claim's core point or supporting evidence.
- helpfulGoodSources should only be selected in addition if the note's correction is fundamentally grounded in clearly cited, authoritative, and faithfully summarized sources (not just the presence of a source).
- helpfulImportantContext is used when the note supplies background or context without which the claim would be misunderstood or misinterpreted, and this context is not a direct factual correction but provides essential interpretive clarity.
- Use helpfulUnbiasedLanguage or helpfulEmpathetic only in conjunction with a substantive clarification or correction and if the neutrality or tone of the explanation is materially helpful. When both direct correction and crucial context are present, prefer (b) helpfulClear or (a) helpfulAddressesClaim as primary, paired with (e) helpfulImportantContext as secondary only if context is indispensable and not redundant with the correction.
- helpfulGoodSources never substitutes for correction—pick it only if the faithful use of sources is a main reason for helpfulness.

3. "NotHelpful" Reason Selection

- Assign "notHelpfulIncorrect" if there is any inaccuracy, factual misstatement, or misleading claim regarding the main point or evidence.
- Assign "notHelpfulMissingKeyPoints" if the note avoids or omits addressing the core issue or supporting fact, no matter how detailed its peripheral info.
- Use "notHelpfulNoteNotNeeded" if the note is trivial, redundant, or supplies information already evident and non-essential for claim comprehension.
- "notHelpfulSourcesMissingOrUnreliable" or "notHelpfulIrrelevantSources" apply if sources are not credible, are misrepresented, or are irrelevant to the core of the claim.
- If tone, personal opinion, bias, speculation, or argumentativeness blocks any clarification, use one of the corresponding "notHelpful" reasons that best fits the limitation.
- For unclear, incomplete, or confusing notes, use "notHelpfulHardToUnderstand." "notHelpfulOther" and "notHelpfulOffTopic" only if none of the above describe the problem or the note is wholly irrelevant.

Reason Definitions:${Reason Definitions}

Decision Process Checklist

1. Is the note factually accurate and not misleading about any major point or evidence?

- If no, assign "notHelpfulIncorrect" and another as fits.
- If yes, proceed.

2. Does the note directly clarify, correct, or critically contextualize the main assertion, a key factual sub-claim, or any central supporting evidence?

- If yes, select the most specific "helpful" reason(s) per above priority order.
- If its primary value is sources, include "helpfulGoodSources" only if the sourcing itself is decisive.
- Do not select "helpfulImportantContext" unless the info is both necessary for accurate interpretation and not primarily a direct correction.
- If note only offers peripheral detail, trivia, or sidesteps the key issue, use "notHelpfulMissingKeyPoints" and/or "notHelpfulNoteNotNeeded."

3. Is the note clear, neutral, and respectful in tone?

- If so in addition to being factually helpful, pair with "helpfulUnbiasedLanguage" or "helpfulEmpathetic" as needed.
- If tone, speculation, or bias prevents meaningful clarification, pick the corresponding "notHelpful" reason.

4. Is the note hard to understand, incomplete, or not addressing the claim?

- Assign "notHelpfulHardToUnderstand" or "notHelpfulOffTopic" as required.

5. Would a typical, reasonably attentive reader gain essential, accurate insight into the claim's truth, context, or credibility—including debunking of misused/incorrect supporting evidence—because of this note?

- If yes, "helpful" reasons most fitting the note's substance.
- If no, most directly explanatory "notHelpful" reasons.

Output in the following JSON format only, no extra output: {"helpfulness": helpful or non_helpful,"reasons":"reason1;reason2"}
CLAIM: ${claim}
NOTE: ${note}
Answer:

Table 13: Optimized English prompt for note helpfulness and reason prediction.

Seed Definitions

Helpful Reasons

- "helpfulClear": "**HelpfulClear** refers to the practice of validating and clarifying claims by providing accurate context or corrections through credible evidence and expert opinion. It involves critically examining information to dispel misinformation, address misconceptions, and ensure that communications are both accurate and informative. The aim is to enhance understanding, promote informed decision-making, and prevent the spread of false or misleading narratives.",
- "helpfulGoodSources": "**HelpfulGoodSources** refers to reliable, fact-based information that clarifies, corrects, or disproves misleading or false claims. It encompasses various forms of authoritative references, including government reports, scientific studies, expert opinions, and credible news articles. The purpose of these sources is to provide accuracy and context to statements made in public discourse, aiding in the dissemination of factual knowledge and supporting informed decision-making.",
- "helpfulAddressesClaim": "**HelpfulAddressesClaim** is a concept that refers to the clarification or correction of claims made in various contexts by providing accurate information, context, or details that counter or explain the original assertion. This often involves highlighting misconceptions, providing factual evidence, or referencing authoritative sources to support the revised understanding of the claim, ultimately helping to prevent misinformation and promoting informed discourse.",
- "helpfulImportantContext": "The REASON, defined as ̈helpfulImportantContext,̈refers to the provision of essential background information that clarifies or corrects claims presented in various statements. This context is crucial for understanding the accuracy, validity, or implications of the claims made, by offering factual corrections, historical context, or relevant comparisons to ensure a more informed and nuanced comprehension of the situation or assertion being discussed.",
- "helpfulInformative": "The definition of the REASON ̈helpfulInformativeïs that it refers to explanations or notes provided in response to claims, which serve to clarify, correct inaccuracies, or provide essential context and background information. These notes help the reader better understand the claims by presenting factual evidence, linking to credible sources, and addressing misunderstandings, thereby enhancing the overall comprehension of the topic at hand.",
- "helpfulUnbiasedLanguage": "The definition of the REASON ̈helpfulUnbiasedLanguageïs the use of clear, neutral, and factual language that provides context and clarification to claims without introducing bias or misinformation. It seeks to present information in an objective manner, ensuring that the audience receives accurate and relevant details to better understand the claims being made. This approach emphasizes transparency and factual integrity, allowing for informed discussions and assessments.",
- "helpfulEmpathetic": "**Helpful Empathetic**: This reason reflects the intention to provide accurate and clarifying information in response to claims that are misleading or false. It embodies a commitment to enhancing understanding by offering factual corrections and context that reveal the truth, thereby enabling informed discussion. The goal is to support others in recognizing discrepancies or inaccuracies in claims, fostering a more informed perspective while acknowledging the importance of factual accuracy."

Continued on next page

Seed Definitions
Unhelpful Reasons
• "notHelpfulOpinionSpeculation": "The definition of the REASON ñotHelpfulOpinionSpeculationïs as follows: It refers to statements or notes that provide personal opinions, unverified claims, or speculative assertions that do not contribute factual information or clarification relevant to the claim being discussed. These opinions or speculative remarks often lack substantiation and do not advance the understanding of the subject matter, thus being non-helpful in the context of verifying or explaining the claims made.",
• "notHelpfulArgumentativeOrBiased": "The REASON "notHelpfulArgumentativeOrBiased" can be defined as follows: This classification refers to comments or notes that fail to provide constructive or relevant context to the accompanying claims. Instead, they tend to be argumentative, inflammatory, or biased, often lacking factual support or logical reasoning. Such responses may attempt to undermine the original claim without addressing its merits, focusing instead on personal attacks, irrelevant comparisons, or emotional appeals that detract from a rational discussion. In essence, they do not foster an understanding of the claim and instead perpetuate division or misinformation.",
• "notHelpfulMissingKeyPoints": "The definition of the REASON ñotHelpfulMissingKeyPointsïs: The notes provided do not adequately address or clarify the claims made. Instead, they often introduce unrelated information or misconstrue the context, leaving significant gaps in understanding the primary arguments presented in the claims.",
• "notHelpfulNoteNotNeeded": "The definition of the REASON ñotHelpfulNoteNotNeededïs that the provided notes fail to contribute meaningful or relevant information to clarify, support, or contradict the claims made. Instead, they either introduce unrelated details, emphasize inaccuracies without addressing the core of the claim, or illustrate a lack of factual backing, thus rendering them ineffective in providing substantiated context or understanding regarding the claims.",
• "notHelpfulIncorrect": "The definition of the REASON ñotHelpfulIncorrectïs as follows: This category refers to claims that are supported by notes or explanations that are factually inaccurate, misleading, or irrelevant to the initial claim. The notes do not effectively clarify, substantiate, or provide accurate context for the claims, thereby failing to contribute meaningful information to the discussion. Instead, they may promote misinformation, misunderstandings, or misinterpretations of the original assertions.",
• "notHelpfulSourcesMissingOrUnreliable": "The REASON for the samples being non-helpful is defined as **the presence of sources that are unreliable, lack credibility, or do not substantiate the claims made**. This includes situations where evidence is anecdotal, opinion-based, or referencing non-verifiable documents, thereby failing to provide the necessary backing to support the claims.",
• "notHelpfulSpamHarassmentOrAbuse": "The definition of the REASON ñotHelpfulSpamHarassmentOrAbuseïs content that is either unconstructive, irrelevant, or disrespectful in nature, often resorting to personal attacks, derogatory language, or baseless accusations. Such comments fail to provide meaningful engagement or contribute to the discussion, instead serving to provoke or demean individuals, typically lacking in factual or supportive content.",
• "notHelpfulIrrelevantSources": "The REASON ñotHelpfulIrrelevantSourcesc̈an be defined as follows: NotHelpfulIrrelevantSources refers to a category of information where the claims being made are accompanied by explanations, notes, or sources that do not provide pertinent or factual support to validate the claims. This includes instances where the sources cited are either misleading, unrelated to the claim, or factually incorrect, resulting in a lack of credible evidence that would otherwise substantiate the assertions made. Essentially, information categorized under NotHelpfulIrrelevantSources fails to enhance understanding or clarify the claims due to its inaccuracy or irrelevance.",
• "notHelpfulOther": "The definition of the REASON ñotHelpfulOtherïs that the provided notes do not effectively contribute to or clarify the claims made. Instead, they often reference unrelated information or distract from the central claim, failing to provide direct support, context, or factual information pertinent to understanding or verifying the claim.",
• "notHelpfulHardToUnderstand": "The REASON for the above samples being non-helpful is that the notes provided do not effectively clarify or support the claims made. They often lack relevant information, context, or clarity, making it difficult for readers to comprehend the claims or understand their significance. This results in a communication gap where the intended message is obscured or misunderstood.",
• "notHelpfulOpinionSpeculationOrBias": "The definition of the REASON ñotHelpfulOpinionSpeculationOrBiasïs as follows: This reason applies to statements or claims that are influenced by subjective views, personal opinions, speculation, or perceived biases rather than presenting objective facts or useful information. Such claims often lack substantiation and do not contribute constructively to the discourse, focusing instead on emotional responses or distorted interpretations of events or behaviors."

Table 14: Note reason seed definitions.

Optimized Definitions
Helpful Reasons
• helpfulClear: The note delivers an accurate, unambiguous correction or direct clarification targeting a central claim or supporting fact/evidence. Use when the note dispels a major misconception, corrects a key number, identifies a misattributed element, or plainly states the essential fact needed to resolve or accurately interpret the claim.
• helpfulAddressesClaim: The note explicitly and accurately responds to or explains the principal assertion, central factual sub-claim, or foundational evidence of the claim. Includes refuting an implied mechanism or performing a core fact check that settles the claim's credibility.
• helpfulGoodSources: The note cites sources that are authoritative, reputable, and faithfully, directly, and unambiguously support its factual correction or clarification. Pick only if sourcing is integral—don't select if other reasons are more fundamental to helpfulness.
• helpfulImportantContext: The note supplies indispensable, targeted background or context needed for the reader to assess the claim or its evidence correctly (such as revealing regulatory nuance, historical policies, deception/fraud, or other context that transforms meaning or evaluation).
• helpfulInformative: The note gives essential factual explanation or contextualization directly linked to dispelling misunderstanding about the claim's main point. Don't use if the note merely provides general or background info.
• helpfulUnbiasedLanguage: The note's language is clear, objective, and neutral while providing a substantive correction or clarification. Use only paired with a direct factual "helpful" label.
• helpfulEmpathetic: The note enhances the factual correction by presenting it in a manner that increases reader trust and comprehension through clarity and respect.
• helpfulUniqueContext: The note delivers information not widely available, central to correctly interpreting the claim or core evidence.
Unhelpful Reasons
• notHelpfulOpinionSpeculation: The note is mainly opinion, speculation, or contains unsubstantiated assertions, not clarifying any core aspect of the claim or its evidence.
• notHelpfulArgumentativeOrBiased: Hostile, inflammatory, or overtly biased language dominates and no factual clarification or relevant context is given.
• notHelpfulMissingKeyPoints: The note wholly fails to clarify, correct, or address the main assertion, supporting fact, or evidence, OR it focuses solely on peripheral, partial, or unrelated points.
• notHelpfulNoteNotNeeded: The note adds no meaningful, fresh, or essential clarification—it is trivial, redundant, tangential, or consists of widely-known facts non-essential for comprehending the claim.
• notHelpfulIncorrect: Contains factual errors, misleading/unsupported claims, or confuses main assertions or evidence.
• notHelpfulSourcesMissingOrUnreliable: Sources are missing, non-credible, don't support the note's claim, or are misrepresented.
• notHelpfulSpamHarassmentOrAbuse: Contains spam, personal attacks, or offensive/unrelated content.
• notHelpfulIrrelevantSources: Sources in the note do not substantiate or clarify any main aspect of the claim or are not relevant to its core.
• notHelpfulOther: Any failure to clarify or support the claim not captured by the above categories.
• notHelpfulHardToUnderstand: The note is so unclear, incomplete, or poorly written that its meaning regarding the claim cannot be determined.
• notHelpfulOpinionSpeculationOrBias: Personal views, speculation, or bias present to such a degree that no clarification on the main claim or evidence is possible.
• notHelpfulOffTopic: The note is wholly unrelated to any claim feature or its main supporting facts/evidence.

Table 15: Optimized reason definitions.

Prompt for Fact Checking Climate Fever claims
Direct Prompt Fact-check the following claim using provided evidence: Claim: ${claim} Evidence: ${evidence_text} Classify the claim as *SUPPORTS*, *REFUTES*, *NOT_ENOUGH_INFO* or *DISPUTED*. Format: Classification: [YOUR_ANSWER] Brief reason: **Incorporating Helpfulness Information** Fact-check this claim using the evidence and the helpfulness information of the evidence, if the evidence is not helpful, take less weight of the evidence. Claim: ${claim} Evidence: ${evidence_text_with_helpfulness_information} Classify the claim as *SUPPORTS*, *REFUTES*, *NOT_ENOUGH_INFO* or *DISPUTED*. Format: Classification: [YOUR_ANSWER] Brief reason:

Table 16: Prompt for Fact Checking `CLIMATE-FEVER` claims.