Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI4Tech: AI Enabling Critical Technologies

Generating Grounded Responses to Counter Misinformation via Learning Efficient Fine-Grained Critiques

Xiaofei Xu , Xiuzhen Zhang * , Ke Deng
School of Computing Technologies, RMIT University, Melbourne, Australia
xiaofei.xu@ieee.org, {xiuzhen.zhang, ke.deng}@rmit.edu.au,

Abstract

Fake news and misinformation poses a significant threat to society, making efficient mitigation essential. However, manual fact-checking is costly and lacks scalability. Large Language Models (LLMs) offer promise in automating counter-response generation to mitigate misinformation, but a critical challenge lies in their tendency to hallucinate non-factual information. Existing models mainly rely on LLM self-feedback to reduce hallucination, but this approach is computationally expensive. In this paper, we propose MisMitiFact, Misinformation Mitigation grounded in Facts, an efficient framework for generating fact-grounded counter-responses at scale. MisMitiFact generates simple critique feedback to refine LLM outputs, ensuring responses are grounded in evidence. We develop lightweight, fine-grained critique models trained on data sourced from readily available fact-checking sites to identify and correct errors in key elements such as numerals, entities, and topics in LLM generations. Experiments show that MisMitiFact generates counter-responses of comparable quality to LLMs' self-feedback while using significantly smaller critique models. Importantly, it achieves ∼5x increase in feedback generation throughput, making it highly suitable for cost-effective, large-scale misinformation mitigation. Code and LLM prompt templates are at https://github.com/xxfwin/MisMitiFact.

1 Introduction

Misinformation spreads faster and farther than truthful information [Vosoughi *et al.*, 2018], posing significant risks to public health, trust, and society. While professional fact-checkers and journalists provide reliable veracity assessments, their efforts are labor-intensive and often focus only on popular claims. Automated fact-checking [Guo *et al.*, 2022; Wang and Shu, 2023; Zeng and Gao, 2024] have been proposed to identify misinformation at scale, but their focus has primarily been on veracity prediction rather than generating direct counter-responses in real-time. The widespread propagation of misinformation, especially on social media platforms, calls for the automated generation of factually grounded counter-responses in real-time and at scale.

The advancement of Large Language models (LLMs) offers the opportunity for automated generation of counter-responses at scale. Leveraging LLMs, [He *et al.*, 2023] proposed to employ LLMs for counter-response generation, but their focus is on the language quality – relevant, fluent, polite and of refutation attitude. However, the critical issue of information factuality, or hallucination, remains largely overlooked. More broadly, studies on the mitigation of LLM hallucination are reported in the literature. Retrieval-Augmented Generation (RAG) approaches augment the input context for LLM generation with retrieved knowledge [Lewis *et al.*, 2020; Yu *et al.*, 2023; Shi *et al.*, 2023] to confine LLM generation and reduce hallucination. Originating from Chain of Thought (CoT) reasoning [Kojima *et al.*, 2022; Zhang *et al.*, 2023], various prompting strategies are proposed to enhance LLM reasoning capabilities to reduce hallucination. Recently, LLM self-feedback strategy [Madaan *et al.*, 2024; Akyürek *et al.*, 2023] is proposed to refine LLM generation, demonstrating strong performance to reduce hallucination. None of these studies directly address the issue of information factuality in LLM outputs. Importantly, all these approaches require additional LLM runs to generate reasoning steps or feedback, which is computationally expensive.

Recent research [Barrón-Cedeno *et al.*, 2020] shows that fact-check-worthy claims typically fall into two categories: potentially harmful claims attacking individuals, organizations, nations, or social groups and time-sensitive claims concerning prominent people, organizations, countries and events. Fact-checking such claims requires precise comparisons of key elements – such as numbers, entities, and topics – against factual evidence. For instance, as shown in Figure 1, a false claim contains key elements such as numbers (e.g. 122,000, Eighty percent), entities (e.g. Australia) and topics (e.g. the problem of homelessness) that must be verified. Here an element refers to an individual piece of information that can be assessed for qualities such as factuality, correctness or relevance.

In this paper, we propose *MisMitiFact* – Misinformation Mitigation grounded in Facts, an efficient framework for gen-

*Corresponding author.

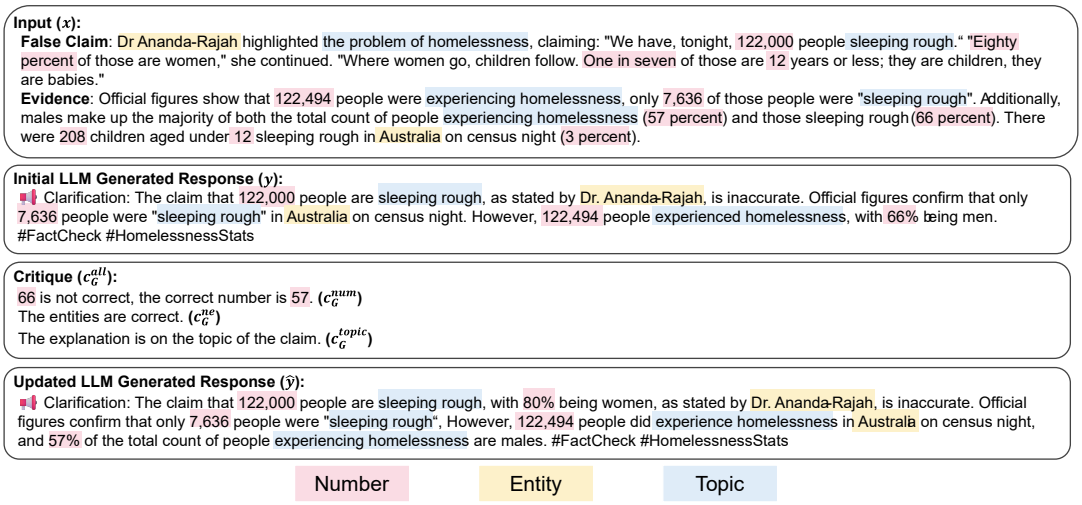

Figure 1: An example for MisMitiFact counter-response generation. Claim and evidence as input, initial LLM generation, the element-based critique feedback prompt on the initial generation, and updated LLM generation. Colours indicate different types of elements.

erating fact-grounded counter-responses to misinformation at scale. Unlike existing approaches that rely on expensive LLM inference for general feedback, our method introduces lightweight, fine-grained critique models that directly pinpoint factual errors about elements such as numbers, named entities, and topics. In our framework, critique models for factual errors about specific elements in the initial generation are trained using data within the readily available fact-checking articles from fact-checking sites. Figure 1 shows how MisMitiFact works. The input x consists of the misinformation claim and the evidence. We collect an initial counter-response, y, from the LLM, then obtain feedback from three critique models, $c_G^{all} = \{c_G^{num}, c_G^{ne}, c_G^{topic}\}$, which respectively evaluate the counter-response on numbers (c_G^{num}), entities (c_G^{ne}), and topics (c_G^{topic}). The LLM then takes the initial counter-response and the critiques into consideration and generates a refined counter-response, $\hat{y}$.

To our best knowledge, we are the first to introduce a task for counter-response generation grounded in factual evidence. We further propose a simple critique feedback approach for LLMs to refine generation and ensure information factuality. We train lightweight, efficient critique models targeting element-wise information correctness, the models are trained on data sourced from readily available fact-checking sites and do not require human-annotated critiques or counter-responses. Experiments on two real-world datasets show that our system MisMitFact can generate grounded counter-responses of quality comparable to the state-of-the-art LLM self-feedback approach while using significantly smaller critique models. Additionally, it achieves ∼5x increase in critique generation throughput, making our system highly suitable for cost-effective, large-scale misinformation mitigation.

2 Related Work

Related work comes from three lines of research.

2.1 Misinformation Mitigation

A line of research on misinformation mitigation focuses on limiting the dissemination of false information while promoting corrective or clarifying truthful information [Farajtabar *et al.*, 2017; Saxena *et al.*, 2020; Goindani and Neville, 2020; Xu *et al.*, 2022; Xu *et al.*, 2024]. These studies typically assume that counter responses containing truthful information are given. Another line of research focuses on automated fact-checking – predicting the veracity of claims based on given claims and evidence [Guo *et al.*, 2022; Wang and Shu, 2023; Zeng and Gao, 2024]. However, their focus is not on generating counter-responses to directly refute false claims.

Some recent studies [He *et al.*, 2023] focus on counter-response generation. Their system MisinfoCorrect employs LLMs to generate counter-responses to COVID-19 vaccine misinformation. While their work models politeness, refutation attitude, and fluency within responses, it does not address the factual correctness of counter-responses, which is a critical issue and the focus of our research.

2.2 Mitigation of LLM Hallucination

A line of research focuses on improving the generation process and broadly includes two classes of studies. One class is based on RAG [Lewis *et al.*, 2020; Yu *et al.*, 2023; Shi *et al.*, 2023], which aims to reduce LLM hallucinations by augmenting the LLM input context with retrieved external knowledge base. Another class leverages CoT prompting [Kojima *et al.*, 2022; Zhang *et al.*, 2023], which improves LLM reasoning capabilities to reduce hallucinations. While these methods enhance the overall quality of LLM generation, they do not directly address the information factuality.

Another line of research provides feedback to refine LLM generation and improve its quality [Schick *et al.*, 2022; Akyürek *et al.*, 2023; Lee *et al.*, 2023; Yu *et al.*, 2023; Madaan *et al.*, 2024]. Early research using textual critique feedback prompts to improve LLM generation, where a separate critique model for generating critiques [Schick *et al.*,

2022; Akyürek *et al.*, 2023] is trained from human-written critiques as training data. A recent study called SELF-REFINE [Madaan *et al.*, 2024] repeatedly calls an LLM to generate feedback on its own generation. This approach not only generates general feedback but also incurs repeated LLM calls, which are computationally expensive. In contrast, our approach introduces lightweight, fine-grained critique models that generate simple critique prompts targeting factual errors in key elements (e.g., numbers, entities, and topics). Our lightweight critique models generate simple, element-wise critiques that incur significantly less computation cost and our MisMitiFact achieves comparable textual quality to LLM-based self-feedback.

2.3 Factual Error Correction

Our general idea of identifying and correcting errors in texts is somewhat related to the NLP task of Factual Error Correction (FEC) for abstractive text summarization [Thorne and Vlachos, 2021; Lee *et al.*, 2022]. FEC focuses on checking the factual correctness of summaries given source documents, a fundamentally different and much simpler problem setting that does not involve claims and their counter-responses.

3 Problem Definition

Our task is to generate counter-response text y to debunk a misinformation claim x_c, given factual evidence text containing facts related to the claim, x_e. Counter-response y is desired to have the following properties:

- *Faithful to the evidence*. Denote all statements in counter-response y as $S = \{s_1, \cdots, s_N\}$ and all verifiable statements in y as $V = \{v_1, \cdots, v_N\}$. The response y is faithful to the evidence iff $\forall s_i \in y, s_i \models x_e$ and $S - V = \emptyset$. For a response to hold this property, all the statements in the response must be verifiable and be verified true ($\models$) according to x_e.
- *Refute the misinformation*. For a successful refutation to the misinformation, the counter-response should satisfy the property of *faithful to the evidence* and also satisfy the following property: Define all false statements in the misinformation claim x_c as $F = \{f_1, \cdots, f_N\}$. Response y refutes the misinformation iff $\forall f_i \in x_c, f_i \not\models y$. For a response to hold this property, all the false statements in the claim must be clarified as false ($\not\models$) by y.

For counter-responses failing to satisfy the property of *faithful to the evidence*, the response may contain non-factual information and is not acceptable when debunking misinformation. For counter-responses failing to satisfy the property of *refute the misinformation*, the responses will not clarify the false claim and thus never counter the misinformation. We aim for a framework that leverages LLMs to automatically generate counter-responses to refute false claims while correcting factual errors contradicting facts in evidence.

4 Methodology

We propose a prompt-learning approach to counter-response generation, focusing on critique feedback to prompt LLMs to correct errors within the initial generation. Inspired by observations that fact-check-worthy claims often contain false information about elements such as numbers, named entities and topics, we propose to train small critique models to pinpoint errors and generate critique feedback for elements (See Figure 1 for an example). We also aim not to incur extra human annotations to train the critique models.

4.1 MisMitiFact

Figure 2 shows our MisMitiFact framework, which can be summarized as follows: (1) A large language model with frozen weights, LLM_{gen}, is used to generate an initial counter-response to a misinformation claim. (2) The initial output of LLM_{gen} is critiqued by a set of three trained small critique models, $LLM_{critique}$. (3) The text output of the critics, along with the initial input to LLM_{gen}, are used to prompt LLM_{gen} to generate a refined output. In contrast to the existing critique-style feedback approach of [Akyürek *et al.*, 2023], which relies on expensive human-written critiques, our approach trains lightweight, critique models on specific elements like numbers, named entities and topics. These models are trained using training data automatically generated from readily available fact-checking articles (Details in Section 4.2).

MisMitiFact consists of two phases: critique model training and counter-response generation. In the critique model training phase, given input x (consists of claim x_c and evidence x_e) and explanation y_E, we generate training data comprising two types of instances to train $LLM_{critique}$: The **factual instances**, where the input x and the original explanation y_E is paired with a positive indicator like "The entities are correct" to train the model to confirm factual accuracy. The **counter-factual instances**, where the factual elements number, name entity and topic in y_E are replaced with false number, false named entity and off-topic, generating y_E^{num}, y_E^{ne}, and y_E^{topic}. Each counter-factual instance is paired with input x and critiques based on templates (c_E^{num}, c_E^{ne}, and c_E^{topic}), indicating the introduced factual errors.

This generated structured training data ensures the $LLM_{critique}$ models learn both to validate factually explanations and to detect counterfactual explanations using critiques. For example, with a false claim x_c =*"122,000 people sleeping rough."* and collected evidence x_e =*"Official figures show that 122,494 people were experiencing homelessness, only 7,636 of those people were sleeping rough."*, a journalist might write an explanation y_E as *"only 7,636 people were sleeping rough."*. We will replace the number *7,636* in y_E with other numbers provided in evidence x_e, that is *122,494*. Thus, the y_E^{num} will be *"only 122,494 people were sleeping rough."* with a corresponding c_E^{num} as *"122,494 is not correct, the correct number is 7,636"*. The same idea also applies to (y_E^{ne}, c_E^{ne}) and $(y_E^{topic}, c_E^{topic})$.

At the inference stage, all models, including the trained critique model $LLM_{critique}$ and the generation model LLM_{gen}, are frozen. When generating the critique for the initially generated counter-response, the critique is given by all the critique models in $LLM_{critique}$ as c_G^{num}, c_G^{ne} and c_G^{topic} respectively. We concatenate all the critiques generated by

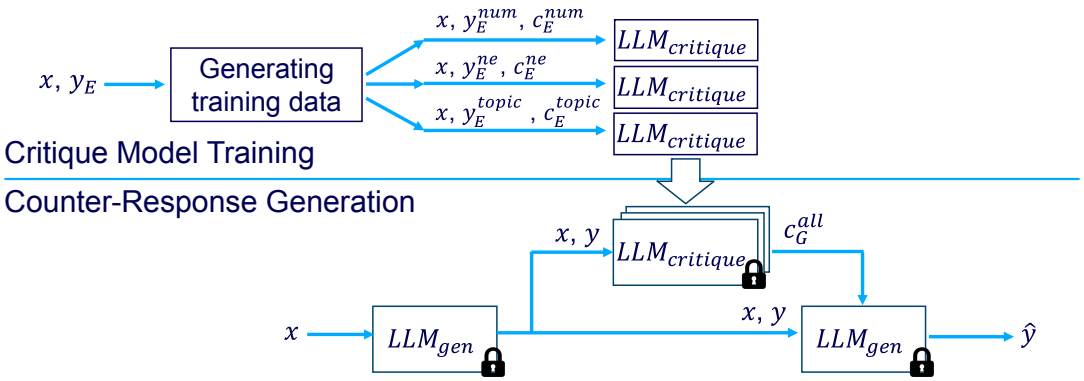

Figure 2: The MisMitiFact framework has two phases: critique model training and counter-response generation. In the critique model training phase, training data are automatically generated to train three $LLM_{critique}$ models on numbers, named entities, and topics. During the counter-response generation phase, a frozen LLM, LLM_{gen}, generates an initial counter-response, which is critiqued by the critique models $LLM_{critique}$, and the critiques are further used as feedback to prompt and refine LLM_{gen} to generate the final counter-responses.

$LLM_{critique}$ as c_G^{all} and feed into the prompt template of LLM_{gen}. Prompt templates of LLM_{gen} are at https://github.com/xxfwin/MisMitiFact.

4.2 Training Data Generation for Critique Models

We next describe how to automatically generate training data for critique models from readily available fact-checking articles. Fact-checking articles usually include four components: (1) a claim, (2) human-curated evidence gathered from credible sources to fact-check the claim, (3) a verdict ("true", "false" or "mixed"), and (4) human explanation to support the fact-check verdict for the claim. As we focus on the false claims, we only include the false claims.

To generate training data for critique models, we focus on using the original claim x_c, the human-curated evidence x_e and the explanation y_E to generate two training components: a) *factual instances*: We pair the explanation y_E with its corresponding claim x_c and evidence x_e, appending affirmative critiques (e.g., "The numbers/entities are correct" or "The explanation is on the topic of the claim"). These instances train the model to confirm factual accuracy when no factual errors exist. b) *counter-factual instances*: The y_E will be modified to generate three types of false explanations - numerically replaced y_E^{num}, named-entity replaced y_E^{ne}, and off-topic y_E^{topic}. For y_E^{num}, we systematically replace number q^y in an explanation y_E with a different number q^x from the evidence text x_e and then create critique c_E^{num} following template "q^x is not correct, the correct number is q^y". Similarly, for y_E^{ne}, we replace named entity q^y in an explanation y_E with a different named entity q^x from the evidence text x_e and then create critique c_E^{ne} following template "q^x is not correct, the correct text is q^y". For y_E^{topic}, we prompt Google Gemini [Team *et al.*, 2023] to generate off-topic messages based on the given evidence. The instruction is to generate off-topic messages that are different from the explanation but still related to the evidence. The training data for the topic critique model is generated by prompting `gemini-pro` using the prompt in Table 1. For each claim x_c, we generate a maximum of 20 instances of numerically replaced and named-entity replaced explanations and generate 3 instances of off-topic explanations. Here, we utilize spaCy [Honnibal and Montani, 2017] to perform named entity recognition to extract all named entities or numbers.

Your task is to rewrite the explanation so that it is off-topic to the provided claim but still on the topic of the provided facts. The output of your response in a single plain json list, please return the rewritten explanation as "rewritten_explanation" and reason as "reason" in the json list. Please also try to keep a similar length of the provided claim. Here is the template of the reason: The claim is about <on topic part of claim>, but the explanation is not correct because it is about <off topic part of explanation>. This is the claim: {claim} This is the explanation: {explanation} This is the facts: {evidence}

Table 1: Prompt for generating training data for topic critique model

5 Experiments

Experiments are conducted on a cluster where each node has 32 cores, 128G memory and is equipped with an NVIDIA Geforce RTX 3090. All deep neural networks are implemented using Transformers [Wolf *et al.*, 2019] under the support of PyTorch [Paszke *et al.*, 2019].

5.1 MisMitiFact and Baselines

For MisMitiFact implementation, we fine-tuned the `T5-large` [Raffel *et al.*, 2020] for the critique models ($LLM_{citiqiue}$ in Fig. 2), with a learning rate of 1e-5, 5

	PUBHEALTH			COVID-19 Vaccine		
Metric	Claim	Evidence	Explanation	Claim	Evidence	Explanation
Avg. No. of Tokens	23.84	1146.69	32.27	53.20	1547.00	37.20
Avg. No. of Numbers	0.51	23.49	0.62	0.49	10.00	0.07
Avg. No. of Entities	0.97	36.04	1.19	1.26	9.00	0.60

Table 2: Statistics of the PUBHEALTH dataset and COVID-19 Vaccine dataset.

epochs, temperature of 1.0 and output length of 5-30 tokens. The overall training time of the three critique models is around 30 hours. Generate critiques are set to maximum 150 tokens. For the generation model of MisMitiFact (LLM_{gen} in Fig. 2) we experimented with two popular open-source LLMs Vicuna-1.5 [Zheng *et al.*, 2024] and LLaMA-2 [Touvron *et al.*, 2023].

We compare MisMitiFact against 5 baseline models. For a fair comparison, we also use the same Vicuna and LLaMA models as the generation models for all baselines.

- *MisinfoCorrect* [He *et al.*, 2023] is a recent model for misinformation counter-response generation model.
- *MisinfoCorrect w/ Evidence* is an extension of MisinfoCorrect, including evidence in the input context.
- *SELF-REFINE* [Madaan *et al.*, 2024] is a recent self-feedback approach that uses an LLM for generation and calls the LLM again to generate feedback and refine its generation. For a fair comparison, we include evidence in the input context.
- *Chain-of-Thought* (CoT) is a popular prompt technique that aims to reduce hallucination in LLMs. We applied zero-shot CoT [Kojima *et al.*, 2022] to instruct the LLMs to solve the task step by step.
- *Plug-and-plug REtrieval FEEDback* (REFEED) [Yu *et al.*, 2023] is a RAG-based model, where the retrieved documents are fed back to the initially generated content, allowing the generation model to refine the content. For our task, the evidence is used as the retrieved documents.

Note that we have not included as baseline models for critic-style feedback to improve LLM generation discussed in Section 2, as they all require gold standard critiques, which are not available in our problem setting.

5.2 Experiment Setup

Datasets: Experiments were conducted on two datasets:

- PUBHEALTH [Kotonya and Toni, 2020] is a large fact-checking dataset containing 11.8K claims on public health topics. For each claim, there is evidence curated by journalists from credible sources, veracity labels (true, false, unproven, mixture) and counter-responses (called "explanations") crafted by journalists. For our experiments, we use the 2153 claims with the "false" or "mixture" label.
- COVID-19 [He *et al.*, 2023] includes false claims about the COVID-19 vaccine from Twitter (now X) and gold standard counter-responses by crowdsourcing. For our experiments, we crawled the source claims based on tweet ID [1] and got 307 false claims. To construct relevant evidence about the claims, we crawled the "Facts about COVID-19 Vaccines" web page from the CDC website [2]. The same evidence on the topic of the COVID-19 vaccine is used for all claims.

We analysed the contents of the claim, evidence and explanation text to understand the characteristic differences between the two datasets. We found that the datasets share a similar count of numbers and entities in the claims, but the counts in the explanations differ a lot. In COVID-19, the explanations contain nearly no numbers. This might be because the majority of false claims can be explained without quoting numbers. From the contents of the evidence, it can be seen that the evidence in PUBHEALTH contains more numbers and named entities, meaning that it is effectively more information-dense than the COVID-19 Vaccine dataset. Table 2 includes statistics of the two datasets used. For all models, we use 80% of claims for training, 10% for development and 10% for testing.

Evaluation Metrics: To evaluate the quality of counter-responses, we employed widely used LLM-based automatic metrics, which are scalable and have demonstrated human-comparable performance [Liu *et al.*, 2023; Min *et al.*, 2023]. We adapted G-EVAL [Liu *et al.*, 2023] based on OpenAI's `GPT-4o-mini` to evaluate four key dimensions: numerical accuracy, named entity accuracy, faithfulness, and refutation. G-EVAL is a framework for evaluating LLM outputs using metrics in a human-like evaluation criterion. We prompt `GPT-4o-mini` to produce scores on a scale of 1 to 5 for each of these dimensions, and scale the score to 0-1 to be consistent with other metrics. We also used FActScore [Min *et al.*, 2023] to assess the factual precision. FActScore evaluates LLM outputs by breaking them down into atomic facts and verifying them against a knowledge source using LLMs. We use `GPT-4o-mini` as the base LLM, and the parameter γ is set as 10 to penalize counter-responses with $<$10 atomic facts. We use the generated counter-responses from the test set to extract atomic facts and use the corresponding evidence as the knowledge source to score the atomic facts. We further compute an Overall score by averaging the scores from G-EVAL and FActScore. This comprehensive metric reflects both the factual accuracy and the effectiveness of the counter-responses in refuting the misinformation.

In practice a false claim may have multiple valid counter-responses focusing on different aspects of the claim and varying in semantic content. We therefore focus on evaluating the factuality – factual consistency with the given evidence – and

[1] https://developer.twitter.com/en/docs/twitter-api

[2] https://www.cdc.gov/coronavirus/2019-ncov/vaccines/facts.html

Model	Numerical↑	Entity↑	Faithfulness↑	Refutation↑	FActScore↑	Overall↑
PUBHEALTH Dataset						
LLM_{gen} = Vicuna						
MisMitiFact (ours)	**0.987**	0.873	0.881	0.716	0.733	0.838
MisinfoCorrect	0.924	0.741	0.661	0.550	0.540	0.683
MisinfoCorrect w/ Evidence	0.908	**0.920**	0.890	**0.777**	0.727	**0.844**
SELF-REFINE	0.822	0.861	0.835	0.652	**0.752**	0.784
CoT	0.673	0.888	**0.892**	0.753	0.719	0.785
REFEED	0.931	0.762	0.668	0.621	0.619	0.720
LLM_{gen} = LLaMA2						
MisMitiFact (ours)	0.889	**0.871**	**0.873**	0.711	0.705	0.810
MisinfoCorrect	0.742	0.672	0.606	0.510	0.495	0.605
MisinfoCorrect w/ Evidence	0.889	0.862	0.854	**0.781**	0.679	0.813
SELF-REFINE	0.917	0.854	0.871	0.751	**0.744**	**0.827**
CoT	**0.967**	0.855	0.859	0.732	0.617	0.806
REFEED	0.733	0.729	0.674	0.574	0.484	0.639
COVID-19 Vaccine Dataset						
LLM_{gen} = Vicuna						
MisMitiFact (ours)	**0.987**	**0.911**	**0.840**	0.690	0.771	**0.840**
MisinfoCorrect	0.931	0.831	0.770	0.745	0.432	0.742
MisinfoCorrect w/ Evidence	0.983	0.862	0.814	**0.763**	0.729	0.830
SELF-REFINE	0.975	0.880	0.789	0.678	**0.777**	0.828
CoT	0.928	0.859	0.820	0.745	0.665	0.803
REFEED	0.756	0.783	0.690	0.705	0.352	0.657
LLM_{gen} = LLaMA2						
MisMitiFact (ours)	0.933	**0.869**	**0.815**	0.717	0.686	0.804
MisinfoCorrect	0.816	0.731	0.630	0.703	0.363	0.649
MisinfoCorrect w/ Evidence	**0.964**	0.829	0.743	0.727	0.490	0.751
SELF-REFINE	0.964	0.863	0.774	**0.731**	**0.697**	**0.806**
CoT	0.619	0.809	0.727	0.685	0.450	0.658
REFEED	0.890	0.769	0.698	0.683	0.399	0.688

Table 3: Experiment results on PUBHEALTH and COVID-19 using LLM-based metrics for factual correctness and refutation

their refutation for the claim. Due to the cost limit of evaluation, we evaluate a total of 100 claims for each dataset, ensuring a representative assessment of our approach.

5.3 Results

Table 3 shows the main experiment result. As can be seen, our MisMitiFact achieves robust performance, particularly in generating refutational counter-responses while maintaining faithfulness. Notably, MisMitiFact consistently matches or surpasses the best-performing baselines in numerical correctness, entity accuracy, faithfulness, and overall refutation quality. On PUBHEALTH, MisMitiFact achieves the highest average score (0.838) when using Vicuna as the LLM_{gen}, outperforming all baselines except for MisinfoCorrect w/ Evidence, which achieves a comparable score of 0.844. When using LLaMA2 as the LLM_{gen}, MisMitiFact maintains strong performance with an average score of 0.810, close to the best-performing baselines like SELF-REFINE (0.827) and MisinfoCorrect w/ Evidence (0.813). On COVID-19, MisMitiFact achieves the highest average score (0.840) with Vicuna as the LLM_{gen}, outperforming all baselines, including SELF-REFINE (0.828) and MisinfoCorrect w/ Evidence (0.830). With LLaMA2 as the LLM_{gen}, MisMitiFact achieves an average score of 0.804, outperforming all baselines except SELF-REFINE, which achieves a comparable score of 0.806. While SELF-REFINE achieves competitive results in terms of FActScore, it is important to note that this approach leverages iterative refinement based on LLM feedback, which may introduce bias in favor of LLM evaluators. In summary, MisMitiFact demonstrates superior or comparable performance against the best-performing baselines across all datasets.

5.4 Accuracy and Efficiency of Critique Models

We apply G-EVAL to evaluate how accurately the critique models pinpoint the inconsistencies between the counter-responses and evidence on a scale of 1 to 5. Table 5 shows the accuracy of the critique models for identifying errors in numbers, entities and topics. It can be seen that overall the critique model shows robust performance, especially when LLM_{gen} = LLaMA2 (with overall ratings of 4.18-4.41). Critique models show the strongest performance on numerical errors (ratings 4.71-4.96 for Vicuna and LLaMA2 on both datasets), but varying performance for topic errors (from ratings of 2.93 for Vicuna on COVID-19 to 4.12 for LLaMA2 on PUBHEALTH). The informal social media language and diverse topics in the COVID-19 dataset may explain this.

We assess the throughput of MisMitiFact critique models on a single machine with a single inference instance to evaluate their efficiency for large-scale applications. Throughput is measured in #critiques/feedbacks per second, a metric that measures how many critique/feedback requests a model can process in one second. MisMitiFact's critique models, based on T5-large (0.738B parameters), achieve a throughput of 0.925 critiques per second. In contrast, SELF-REFINE, which uses LLaMA2 (6.74B parameters) for feedback gener-

Model	Numerical↑	Entity↑	Faithfulness↑	Refutation↑	FActScore↑	Overall↑
PUBHEALTH Dataset						
LLM_{gen} = Vicuna						
MisMitiFact (ours)	**0.987**	0.873	**0.881**	**0.716**	0.733	**0.838**
MisMitiFact w/o NNE	0.926	**0.879**	0.878	0.695	**0.756**	0.827
MisMitiFact w/o T	0.897	0.844	0.825	0.672	0.750	0.798
LLM_{gen} = LLaMA2						
MisMitiFact (ours)	0.889	0.871	0.873	0.711	0.705	0.810
MisMitiFact w/o NNE	0.856	**0.872**	**0.890**	**0.742**	**0.708**	**0.813**
MisMitiFact w/o T	**0.922**	0.842	0.835	0.714	0.626	0.788
COVID-19 Vaccine Dataset						
LLM_{gen} = Vicuna						
MisMitiFact (ours)	0.987	**0.911**	**0.840**	**0.690**	0.771	**0.840**
MisMitiFact w/o NNE	**0.988**	**0.911**	0.815	**0.690**	**0.774**	0.836
MisMitiFact w/o T	0.980	0.872	0.775	0.683	0.770	0.816
LLM_{gen} = LLaMA2						
MisMitiFact (ours)	0.933	**0.869**	**0.815**	0.717	0.686	**0.804**
MisMitiFact w/o NNE	0.955	0.866	0.796	**0.721**	0.674	0.802
MisMitiFact w/o T	**0.972**	0.855	0.773	0.692	**0.707**	0.800

Table 4: The ablation study result for MisMitiFact

	Numerical↑	Entity↑	Topic↑	Overall↑
PUBHEALTH				
LLM_{gen}=Vicuna	4.77	4.44	3.86	4.21
LLM_{gen}=LLaMA2	4.71	4.48	4.12	4.41
COVID-19				
LLM_{gen}=Vicuna	4.98	4.33	2.93	3.36
LLM_{gen}=LLaMA2	4.96	4.38	4.08	4.18

Table 5: Accuracy (5-point scale) of the critique models

ation, achieves a throughput of only 0.165 critiques per second. MisMitiFact achieved a 5.6 times (0.925 vs. 0.165) improvement in critique/feedback throughput, and this significant difference in throughput highlights the efficiency of MisMitiFact. MisMitiFact critique models are not only significantly smaller in size (0.738B vs. 6.74B parameters) but also faster in processing critiques, making them more suitable for large-scale misinformation mitigation.

The high throughput of MisMitiFact can also be attributed to the simple, short critiques generated by MisMitiFact in contrast to the long narrative feedback generated by SELF-REFINE. As shown in the example in Fig. 1, MisMitiFact generates critiques on numbers, facts and topics – in 3 sentences of a total of 28 tokens. In contrast, SELF-REFINE generates a feedback narrative like "Thank you for providing the claim and the facts. Here's my feedback on the original explanation: ... the explanation does not address the fact that males make up the majority of both the total count of people experiencing homelessness (57%) and those sleeping rough (66%)." – in 5 sentences of a total of 124 tokens.

5.5 Ablation Study

We evaluated several variations of MisMitiFact with different parts of critique models:

- *MisMitiFact w/o number and named entity critique models (NNE)* is the variation that removes the number and named entity critique models and only includes the topic critique model.
- *MisMitiFact w/o topic critique model (T)* is the variation that removes the topic critique model and only includes the number and named entity critique models. Number and named entity critique models are treated together since they both contribute to the faithfulness of the generated content.

Table 4 shows the experiment result of the ablation study. We use the same evaluation metrics as in Table 3. It can be observed that MisMitiFact has the best performance on the overall performance in nearly all settings, the performance gain is achieved by both types of critiques. Removing NNE degrades numerical accuracy and faithfulness in specific settings (e.g., Vicuna on PUBHEALTH), while removing T degrades faithfulness (e.g., COVID-19 dataset). In FActScore, MisMitiFact has comparable performance to the best-performing settings, demonstrating the robust ability to generate counter-responses faithful to the evidence.

6 Conclusion

This study tackled the challenge of generating grounded responses to counter misinformation. We proposed to generate simple critique feedback for LLMs to refine their initial generation and ensure responses are grounded in evidence. Our MisMitFact framework trains lightweight critique models on data sourced from readily available fact-checking sites to pinpoint errors in key elements like numbers, named entities, and topics in LLM generations. Experimental results show that MisMitiFact can generate counter-responses of comparable quality to LLMs' self-feedback while using significantly smaller critique models. Additionally, it achieves ∼5x increase in feedback generation throughput, making it highly suitable for cost-effective, large-scale misinformation mitigation. Our future work will focus on counter-response generation in the presence of noisy evidence.

Acknowledgments

This research is supported in part by the ARC Discovery Projects DP200101441, DP210100743 and ARC Linkage Project LP180100750.

References

[Akyürek *et al.*, 2023] Afra Feyza Akyürek, Ekin Akyürek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, 2023.

[Barrón-Cedeno *et al.*, 2020] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 215–236. Springer, 2020.

[Farajtabar *et al.*, 2017] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. Fake news mitigation via point process based intervention. In *International Conference on Machine Learning*, pages 1097–1106. PMLR, 2017.

[Goindani and Neville, 2020] Mahak Goindani and Jennifer Neville. Social reinforcement learning to combat fake news spread. In *Uncertainty in Artificial Intelligence*, pages 1006–1016. PMLR, 2020.

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

[He *et al.*, 2023] Bing He, Mustaque Ahamad, and Srijan Kumar. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709, 2023.

[Honnibal and Montani, 2017] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[Kotonya and Toni, 2020] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, 2020.

[Lee *et al.*, 2022] Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim, and Kyomin Jung. Factual error correction for abstractive summaries using entity retrieval. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 439–444, 2022.

[Lee *et al.*, 2023] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[Liu *et al.*, 2023] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, 2023.

[Madaan *et al.*, 2024] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[Min *et al.*, 2023] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[Saxena *et al.*, 2020] Akrati Saxena, Wynne Hsu, Mong Li Lee, Hai Leong Chieu, Lynette Ng, and Loo Nin Teow. Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. In *Companion Proceedings of the Web Conference 2020*, pages 363–370, 2020.

[Schick *et al.*, 2022] Timo Schick, A Yu Jane, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave,

and Sebastian Riedel. Peer: A collaborative language model. In *The Eleventh International Conference on Learning Representations*, 2022.

[Shi *et al.*, 2023] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Thorne and Vlachos, 2021] James Thorne and Andreas Vlachos. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309, 2021.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[Wang and Shu, 2023] Haoran Wang and Kai Shu. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, 2023.

[Wolf *et al.*, 2019] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[Xu *et al.*, 2022] Xiaofei Xu, Ke Deng, and Xiuzhen Zhang. Identifying cost-effective debunkers for multi-stage fake news mitigation campaigns. In *International Conference on Web Search and Data Mining*, 2022.

[Xu *et al.*, 2024] Xiaofei Xu, Ke Deng, Michael Dann, and Xiuzhen Zhang. Harnessing network effect for fake news mitigation: Selecting debunkers via self-imitation learning. *arXiv preprint arXiv:2402.03357*, 2024.

[Yu *et al.*, 2023] Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*, 2023.

[Zeng and Gao, 2024] Fengzhu Zeng and Wei Gao. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *arXiv preprint arXiv:2401.08026*, 2024.

[Zhang *et al.*, 2023] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

[Zheng *et al.*, 2024] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.