Evidence-Driven Retrieval Augmented Response Generation for Online Misinformation

Zhenrui Yue Huimin Zeng Yimeng Lu Lanyu Shang
Yang Zhang Dong Wang
University of Illinois Urbana-Champaign
{zhenrui3, huiminz3, yimengl4, lshang3, yzhangnd, dwang24}@illinois.edu

Abstract

The proliferation of online misinformation has posed significant threats to public interest. While numerous online users actively participate in the combat against misinformation, many of such responses can be characterized by the lack of politeness and supporting facts. As a solution, text generation approaches are proposed to automatically produce counter-misinformation responses. Nevertheless, existing methods are often trained end-to-end without leveraging external knowledge, resulting in subpar text quality and excessively repetitive responses. In this paper, we propose retrieval augmented response generation for online misinformation (RARG), which collects supporting evidence from scientific sources and generates counter-misinformation responses based on the evidences. In particular, our RARG consists of two stages: (1) evidence collection, where we design a retrieval pipeline to retrieve and rerank evidence documents using a database comprising over 1M academic articles; (2) response generation, in which we align large language models (LLMs) to generate evidence-based responses via reinforcement learning from human feedback (RLHF). We propose a reward function to maximize the utilization of the retrieved evidence while maintaining the quality of the generated text, which yields polite and factual responses that clearly refutes misinformation. To demonstrate the effectiveness of our method, we study the case of COVID-19 and perform extensive experiments with both in- and cross-domain datasets, where RARG consistently outperforms baselines by generating high-quality counter-misinformation responses.

1 Introduction

As social media discussions on trending topics continue to grow, increased presence of online misinformation has been observed on such sources (Chen et al., 2022; Chen and Shu, 2023b). Yet timely intervention against the spreading of misinformation is often lacking, resulting in potential threats to public interest (Ball and Maxmen, 2020; Roozenbeek et al., 2020). For instance, vaccine hesitancy is strongly associated with the exposure of anti-vaccine misinformation (Pierri et al., 2022). Therefore, it is essential to identify and curb the spreading of misinformation before it leads to severe consequences (Litou et al., 2017; Zhu et al., 2021). To identify text-based misinformation, various approaches have been proposed by leveraging language models (Shu et al., 2020; Wu et al., 2022c; Shu et al., 2022; Yue et al., 2022, 2023; Chen and Shu, 2023a; Liu et al., 2024). Upon the detection of misinformation, crowdsourcing approaches (e.g, users, fact-checkers) are designed to prevent the continued propagation of misinformation via counter-responses (Veeriah, 2021; Vraga and Bode, 2021; Kou et al., 2022a; Seo et al., 2022; Kou et al., 2022b; Bozarth et al., 2023; Drolsbach and Pröllochs, 2023). Nevertheless, such responses to misinformation often lack politeness, direct refutation or fail to provide supporting evidence (He et al.,

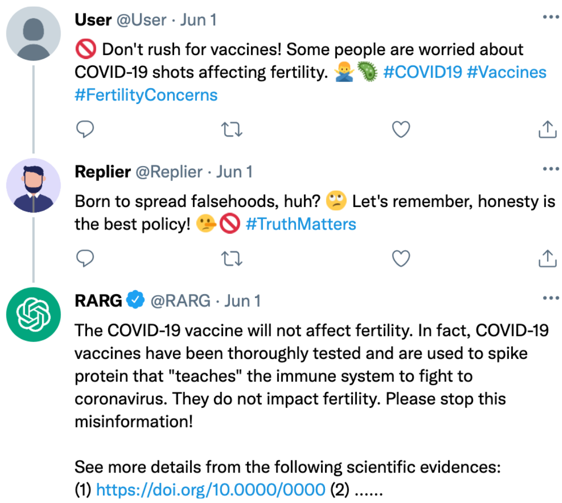

Figure 1: The proposed retrieval augmented response generation (RARG), which collects scientific evidence to generate counter-misinformation responses.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5628–5643
June 16-21, 2024 ©2024 Association for Computational Linguistics

2023; Tafur and Sarkar, 2023). Moreover, human efforts are often less effective when confronted with an overwhelming amount of online misinformation (Micallef et al., 2020, 2022).

To facilitate early intervention at scale, language generation methods are proposed to generate counter-responses against detected misinformation (Vo and Lee, 2019, 2020; He et al., 2023). For example, MisinfoCorrect adopts reinforcement learning to generate responses to refute misinformation (He et al., 2023). Despite their effectiveness, existing generation methods are trained end-to-end on a specific domain. As a result, the generated texts often highly resemble the training responses and demonstrate deteriorated quality upon domain shifts (as we show in Section 4). Additionally, such generation models are unaware of external knowledge and must be frequently updated to incorporate up-to-date domain knowledge (Lewis et al., 2020b; Izacard and Grave, 2021; Borgeaud et al., 2022; Asai et al., 2023). As such, we consider a retrieval augmented generation (RAG) setting that incorporates relevant scientific documents to utilize external evidence without retraining. That is, given input misinformation, our first objective is to retrieve evidence from an extensive collection of documents (e.g., academic publications). Then, we select the relevant documents as evidence to perform retrieval augmented generation, with the goal of extracting supporting facts to debunk misinformation. We illustrate an example of our framework in Figure 1, where evidence-backed counter-response is generated upon the identification of misinformation.

In this paper, we focus on counter-response generation against online misinformation and propose an evidence-driven retrieval augmented response generation (RARG) framework, which efficiently retrieves supporting documents and generates responses using the collected supporting facts. Specifically, RARG consists of two modules: (1) evidence collection, in which we design a two-stage retrieval pipeline based on a collection of over 1M academic articles we collected. For efficient retrieval, our pipeline first performs a coarse search over the data collection, followed by a reranking stage with improved relevance estimation of the retrieved documents; (2) evidence-based response generation, here, we align large language models (LLMs) to generate responses upon the collected evidence via reinforcement learning from human feedback (RLHF). In particular, we propose a reward design that maximizes the utilization of the retrieved evidence while maintaining the quality of the responses. As such, our RARG generates polite and factual responses that clearly refutes input misinformation. To the best of our knowledge, we are the first to introduce a retrieval augmented generation framework with two-stage retrieval and fine-grained RLHF. To demonstrate the effectiveness of RARG, we study the case of COVID-19, where we perform both in- and cross-domain experiments with comprehensive quantitative and qualitative analyses. Experimental results highlight the effectiveness of the proposed framework, where RARG consistently outperforms state-of-the-art baselines by generating high-quality evidence-based responses against online misinformation.

We summarize our contributions as follows:

1. We propose a retrieval augmented generation setting for counter-misinformation response generation. In this work, we focus on COVID-19 misinformation and collected over 1M academic publications as the source for evidence-based response generation.

2. We design RARG, a response generation framework against online misinformation. Our framework combines two-stage retrieval for evidence collection and RLHF-based LLM alignment, such that RARG is optimized to generate polite and factual counter-responses.

3. We show the effectiveness of RARG by experimenting on both in-domain and cross-domain COVID misinformation to validate the generalization of RARG. Both quantitative and qualitative results demonstrate that RARG can outperform state-of-the-art methods in generating high-quality responses.

2 Related Work

2.1 Detecting and Countering Misinformation

Existing methods for detecting misinformation can be broadly categorized into: (1) content-based detection, in which machine learning models are trained upon input contents (i.e., claims) to perform classification (Yue et al., 2022; Zeng et al., 2022; Jiang et al., 2022; Yue et al., 2023; Chen and Shu, 2023a; Liu et al., 2023; Mendes et al., 2023; Zeng et al., 2024; Liu et al., 2024). Additional modalities such as image, video or propagation paths can also be used to improve detection performance (Shang et al., 2021; Santhosh et al., 2022; Shang et al., 2022b; Wu et al., 2022c; Zhou et al., 2023); (2) evidence-based detection, where external knowledge can be collected as supporting evidence to verify input contents (Brand et al., 2021; Kou et al., 2022a; Wu et al., 2022a; Yang et al., 2022; Shang et al., 2022c; Xu et al., 2022; Zhao et al., 2023). For example, knowledge graphs or retrieved document pieces can be processed to support or refute statements (Kou et al., 2021; Hu et al., 2021b; Koloski et al., 2022; Kou et al., 2022b; Shang et al., 2022a; Wu et al., 2022b).

To curb the spreading of misinformation, the majority of existing approaches relies on social correction (e.g., comments, replies) by users or experts (Veeriah, 2021; Vraga and Bode, 2021; Seo et al., 2022; Bozarth et al., 2023). Despite their effectiveness, many of such responses can be characterized by the lack of politeness and supporting facts. Furthermore, user efforts often fall short when confronted with an overwhelming amount of misinformation (Micallef et al., 2020, 2022; He et al., 2023). Recently, language-based methods are proposed to generate counter-responses / explanations (Vo and Lee, 2019, 2020; He et al., 2023; Wan et al., 2024). However, these methods concentrate on improving text quality, often neglecting the response factuality through the incorporation of evidence. Hence, we aim to design a counter-misinformation response generation framework via the collection of relevant documents, followed by the reasoning and generation of responses.

2.2 Retrieval Augmented Generation

Recent progress in large language models (LLMs) has led to substantial improvements in both language understanding and generation (Raffel et al., 2020; Brown et al., 2020; Wei et al., 2021; Ouyang et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; OpenAI, 2023). Thanks to the extensive corpora used during pretraining, LLMs have the ability to embed world knowledge in their parameters, and thus achieve significant performance enhancements across various scenarios (Touvron et al., 2023; OpenAI, 2023; Penedo et al., 2023). Nevertheless, LLMs struggle to capture fine-grained knowledge and frequently exhibit instances of hallucination (Sun et al., 2023; Peng et al., 2023). To incorporate up-to-date knowledge without expensive retraining, retrieval augmented generation is proposed to generate text conditioned on collected documents (Lewis et al., 2020b; Izacard and Grave, 2021; Borgeaud et al., 2022; Izacard et al., 2022; Shi et al., 2023; Ram et al., 2023). For example, REALM retrieves from a large set of documents such as Wikipedia to solve conditional generation tasks like question answering (Guu et al., 2020). Nevertheless, current retrieval augmented methods primarily study knowledge-intensive NLP tasks and have not been well researched for response generation against online misinformation, let alone the combination of fine-grained retrieval with RLHF-based alignment. As such, our work explores retrieval augmented response generation by collecting scientific evidence and performing RLHF alignment to generate evidence-based counter-responses.

3 Methodology

3.1 Preliminary

We consider the following problem setup for counter-misinformation response generation: given input misleading claim x, our objective is to: (1) collect a set of m scientific evidence $\{e_i\}_{i=1}^{m}$ that are relevant to input x to be used as supporting evidence; and (2) generate response y upon input x and the retrieved evidence $\{e_i\}_{i=1}^{m}$, which should demonstrate certain desirable properties (e.g., politeness and factuality), see example in Figure 1. We elaborate our framework in the following.

Input & Output: We denote the evidence collection model as f_{ret} and the response generation model as f_{gen}. Formally, our research framework can be defined as two sub-problems of information retrieval (i.e., evidence retrieval) and retrieval augmented text generation (i.e., response generation), with each of the settings defined as follows:

- *Evidence Retrieval*: Given input misinformation x, human annotated evidence e and a collection of n evidence documents $\{e_i\}_{i=1}^{n}$ (with $e \in \{e_i\}_{i=1}^{n}$), the model f_{ret} should ideally generate the highest relevance score for the claim-evidence pair (i.e., $f_{\text{ret}}(x, e) = \max\{f_{\text{ret}}(x, e_i)\}_{i=1}^{n}$). During training, input x and e can be used to optimize f_{ret}. In inference, we collect a subset of m documents $\{e_i\}_{i=1}^{m}$ for response generation, with $m \ll n$.

- *Response Generation*: For response generation purpose, we incorporate both input claim x and collected evidence subset $\{e_i\}_{i=1}^{m}$ from the previous step in the input prompt, while y represents the generated response from f_{gen} (i.e., $y = f_{\text{gen}}(x, \{e_i\}_{i=1}^{m})$). To learn the generation model, we first train f_{gen} end-to-end, followed RLHF tuning. The inference stage is completed by collecting evidence documents and performing evidence-based response generation.

Optimization: The retrieval model f_{ret} and generation model f_{gen} are parameterized by θ_{ret} and θ_{gen}. To learn f_{ret}, we maximize the relevance score of the label evidence e upon input x. In other words, we minimize the expected loss of input-evidence pair (x, e): $\min_{\theta_{\text{ret}}} \mathbb{E}_{(x,e)\sim\mathcal{X}_{\text{ret}}}[\mathcal{L}(\theta_{\text{ret}}, (x, e))]$, with $\mathcal{X}_{\text{ret}}$ representing the evidence retrieval dataset. For f_{gen}, the optimization is two-fold: (1) θ_{gen} is first trained end-to-end by minimizing cross entropy loss on input-evidence-response triplets, where m can be empirically selected (i.e., $\min_{\theta_{\text{gen}}} \mathbb{E}_{(x,\{e_i\}_{i=1}^m,y)\sim\mathcal{X}_{\text{gen}}}[\mathcal{L}(\theta_{\text{gen}}, (x, \{e_i\}_{i=1}^m, y))]$); (2) to improve the response quality, we leverage an additional reinforcement learning step to tune θ_{gen} towards generating responses of human preference. In particular, we design the reward by considering the following aspects that could help reducing the spread of misinformation (Starbird et al., 2014; Chan et al., 2017; Tanaka and Hirayama, 2019; Malhotra et al., 2022; He et al., 2023):

- *Refutation*: Refutation involves phrases within the response that explicitly and objectively expose the error or inaccuracy of input claims.
- *Factuality*: Factuality evaluates the correctness and supporting evidence of the response, which reflects the reliability of the generated text.
- *Politeness*: Politeness refers to using considerate and thoughtful language, and thus respectfully counters misinformation and avoids backfire.
- *Claim Relevance*: Response should establish a clear and direct connection to the input claim for improved coherence and comprehension.
- *Evidence Relevance*: The relevant evidence information should be included in the response to demonstrate the falsehood of input claim.

To summarize, we introduce a retrieval augmented response generation framework with two-stage evidence retrieval and fine-grained reinforcement learning, which learns to construct evidence-based counter-responses against misinformation.

3.2 Two-Stage Evidence Retrieval

Existing retrieval augmented generation approaches adopts unsupervised sparse retrieval to collect relevant documents (Lewis et al., 2020b; Izacard and Grave, 2021; Izacard et al., 2022; Ram et al., 2023). An example of such algorithms include BM25, which computes query-document relevance by considering word frequency in both query and documents (Robertson et al., 1995). Although BM25 is widely used, it does not yield satisfactory ranking performance for knowledge-intensive tasks such as question answering. Therefore, dense retrieval methods are proposed for improved text understanding and relevance estimation (Karpukhin et al., 2020; Ren et al., 2021; Izacard et al., 2021; Wang et al., 2022, 2023b). Nevertheless, dense retrieval is known to be inefficient for massive data quantity and require large amounts of annotated data for training. This limitation renders dense retrieval to be less effective in retrieval augmented response generation, where only limited annotated data is available for training.

Unlike existing retrieval methods, we design a two-stage retrieval pipeline in RARG that performs coarse-to-fine ranking, which improves the computation efficiency and retrieval performance with limited data. Specifically, we include the following stages: (1) retrieve a smaller subset from the large collection of evidence documents via BM25; and (2) rerank the retrieved subset using a dense retriever fine-tuned on limited claim-evidence pairs. In the first retrieval stage, we adopt BM25 to efficiently generate a small subset $\{e_i\}_{i=1}^m$ from a much large collection $\{e_i\}_{i=1}^n$. While BM25 may occasionally retrieve less or ir-relevant documents, we observe that in most cases, the desired evidence can still be found in $\{e_i\}_{i=1}^m$ with proper selection of m. In our implementation, we adopt $m = 20$ as the subset size to balance the performance and efficiency of our retrieval pipeline. In the following stage, we fine-tune a dense retrieval model to perform fine-grained reranking and select the most relevant documents as supporting evidence. Compared to sparse retrieval, this additional stage yields notable performance improvements with minimal computational overhead. We illustrate the overall retrieval pipeline of RARG in Figure 2 (left).

For dense retriever f_{den} within f_{ret}, fine-tuning often requires large amounts of annotated data for improved ranking performance and generalizability (Karpukhin et al., 2020; Wang et al., 2022). Despite the insufficiency of data, the retrieval results from our first stage can be used as a coarse estimation of claim-evidence relevance. In other words, it is possible to leverage the BM25 scores

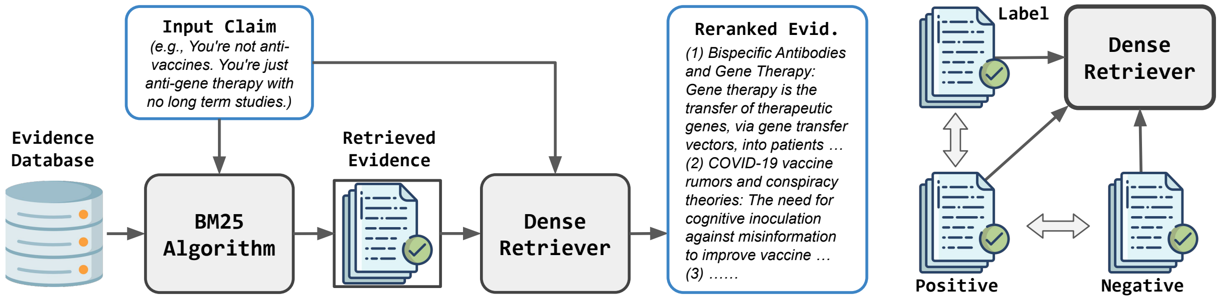

Figure 2: The proposed two-stage retrieval pipeline in RARG (left) and its optimization (right).

from the first stage, combined with limited annotated examples, to optimize the dense retrieval model. Specifically for input-evidence pair (x, e), we sample k positive documents $\{e_i^p\}_{i=1}^k$ and k negative documents $\{e_i^n\}_{i=1}^k$ using BM25 and inverse BM25 scores, which avoids introducing extensive noise in the training data. Based on the positive set $\{e_i^p\}_{i=1}^k$, we design a ranking loss to enlarge the margin between evidence e and the highest ranked positive evidence in $\{e_i^p\}_{i=1}^k$ (i.e., $f_{\text{den}}(x, e) - \max\{f_{\text{den}}(x, e_i^p)\}_{i=1}^k)$). Here, we maximize the relevance between input-evidence pairs by applying penalty only when the margin is below threshold τ. In addition to the ranking loss, our optimization goal contains a contrastive term based on InfoNCE loss (Chen et al., 2020), which improves the relevance estimation between input-evidence pairs while 'pushing away' negative evidence. Formally, the overall optimization objective $\mathcal{L}$ for f_{den} can be formulated as:

$$\mathbb{E}_{(x,e)\sim\mathcal{X}_{\text{ret}}}[\max(0, \max(\{f_{\text{den}}(x, e_i^p)\}_{i=1}^k) - f_{\text{den}}(x, e) + \tau) \\ - \lambda \frac{\exp(f_{\text{den}}(x, e))}{\exp(f_{\text{den}}(x, e)) + \sum_k \exp(f_{\text{den}}(x, e_i^n))}], \tag{1}$$

where τ is the margin threshold for the ranking loss and λ is a scaling factor. For each pair of x and e, the first term in Equation (1) is effective when $f_{\text{den}}(x, e)$ score is not τ greater than the score of the highest ranked positive evidence. The next term, on the other hand, maximizes exponential score of the annotated input-evidence pair over the sum of those from the sampled negative evidence. We illustrate the optimization of the dense retriever in Figure 2 (right), where we differentiate the annotated evidence e from the positive evidence $\{e_i^p\}_{i=1}^k$ and 'push away' the negative evidence $\{e_i^n\}_{i=1}^k$.

For the evidence document collection, since we focus on the case of COVID-19, we limit the evidence documents to academic articles on COVID research from reliable sources. In particular, we collect all available articles from two large databases of COVID-related research: CORD (Wang et al., 2020) and LitCovid (Chen et al., 2021). In sum, we obtain 1,056,262 articles from CORD and 301,136 articles from LitCovid[1]. The collected articles are preprocessed to filter invalid and repeated items. As a result, 1,118,112 articles remain in our evidence database and we extract title and abstract information from each article as the evidence corpus. To fine-tune our dense retriever model, we utilize Check-COVID, a dataset with ∼350 evidence documents and ∼1k input-evidence pairs for training, with all documents collected from the CORD dataset (Wang et al., 2023a). To improve the ranking performance over large evidence collections, we manually increase the document size of Check-COVID to 5k by sampling additional documents from CORD and LitCovid. We use $k = 4$ in $\{e_i^p\}_{i=1}^k$ and $\{e_i^n\}_{i=1}^k$, we also provide more data collection and processing details in Section 4 and Appendix A.

3.3 Response Generation

Despite recent progress in large language models (LLMs), it is frequently observed that LLMs demonstrate hallucination behavior due to the lack of knowledge and reasoning capabilities (Sun et al., 2023; Peng et al., 2023). To improve the factuality of LLM-generated text, retrieval augmented text generation is used to incorperate external knowledge (Lewis et al., 2020b; Izacard and Grave, 2021; Borgeaud et al., 2022; Ram et al., 2023). Yet the training of such models heavily relies on the quality of the data and demands extensive tuning efforts to achieve substantial performance improvements. On the other hand, generating large amounts of responses with desired properties (e.g., factuality,

[1]CORD is maintained until June 2022, while LitCovid is active as of October 2023. Our LitCovid collection is updated with information available until August 2023.

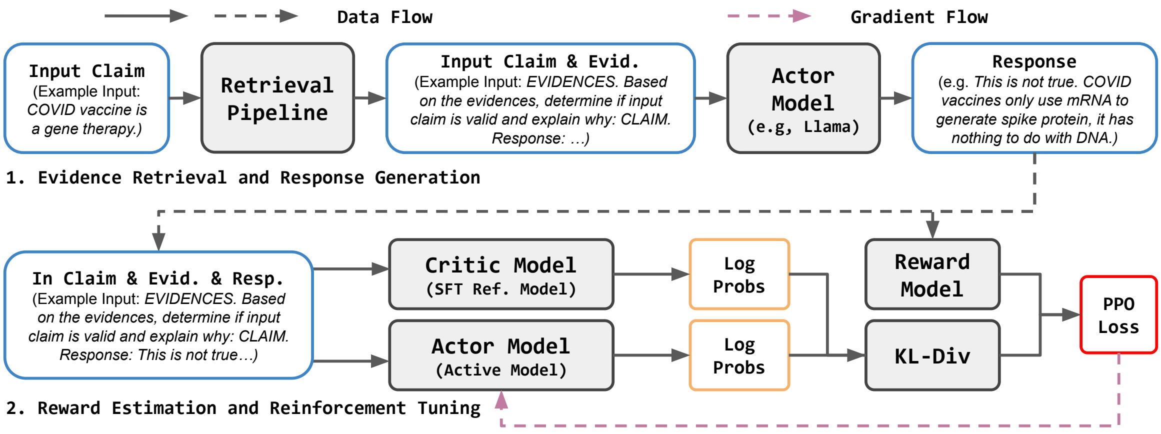

Figure 3: The optimization of f_{gen} in RARG. Upon input claim, the evidence documents are retrieved for response generation. Then, the reward model estimates the rewards and update the actor model with PPO-based RL.

politeness) can be time-consuming and costly. In contrast, gathering basic human feedback such as preference ranking proves to be a notably simpler and more cost-effective approach for feedback collection (Ouyang et al., 2022). Therefore, we combine the advantages of LLMs, which can be easily instruction-tuned with small-size data, while exploiting the potential for further improvement through RLHF (Ouyang et al., 2022).

To optimize the generation model, we propose a RLHF-based approach to generate responses towards human preference. Specifically, f_{gen} is first fine-tuned in a supervised fashion to generate counter-response conditioned on input claim and retrieved evidence. At the same time, a reward model is constructed by leveraging binary human feedback (e.g., whether responses are refuting, factual and polite) as training data to evaluate the response quality. The reward model is then applied in the reinforcement learning stage, where the fine-tuned f_{gen} (i.e., actor model) is trained to further improve generation quality using proximal policy optimization (PPO) (Schulman et al., 2017). Upon deployment, only the actor model is required for inference, which generates counter-misinformation responses based on input claims and collected evidence. We provide an illustration for the response generation process and its optimization in Figure 3.

Supervised Fine-Tuning. Reinforcement learning is known for being unstable for complex tasks, therefore, we first perform supervised fine-tuning using a pretrained LLM (e.g., Llama 2) to learn initial capabilities in counter-misinformation response generation. Additionally, the fine-tuned model is used as a reference model (i.e., critic model in Figure 3) in the following reinforcement learning step to stabilize training. We denote the supervised fine-tuned model as π_{ref} (i.e. reference model).

Modeling Reward. The purpose of reward modeling is to leverage human feedback for evaluating the text quality. Unlike methods that generate text pairs for preference evaluation (Ouyang et al., 2022), we train smaller LMs to model human feedback, as the swift convergence accelerates reward estimation in RLHF. For a dataset of $\{x_i, y_i, r_i\}_{i=1}^n$ with $r \in \{0, 1\}$ representing the reward label from human feedback (i.e., refutation, factuality or politeness in binary form), the objective of modeling reward is to learn a modeling function f that minimizes the expected negative log-likelihood loss. The learnt reward model evaluates the quality of the generated responses. In our implementation, we adopt three BERT models to learn human feedback in refutation, factuality, politeness, and use the sum of the scores as reward. To further improve the coherence of the response w.r.t. input claim and retrieved evidence, we assess the relevance between claim-response pairs (i.e, (x, y)) and evidence-response pairs (i.e, (e, y)). To maximize parameter efficiency, we utilize f_{den} from the previous retrieval module, since f_{den} is specifically trained to model the evidence relevance. Thus the estimated reward $\hat{r}$ is formulated as:

$$\begin{aligned}\hat{r} = & f_{\text{refutation}}(x, \{e_i\}_{i=1}^m, y) + f_{\text{factuality}}(x, \{e_i\}_{i=1}^m, y) \\ & + f_{\text{politeness}}(x, \{e_i\}_{i=1}^m, y) + \alpha(f_{\text{den}}(x, y) \\ & + \max\{f_{\text{den}}(e_i, y)\}_{i=1}^m),\end{aligned} \tag{2}$$

where $f_{\text{refutation}}$, $f_{\text{factuality}}$ and $f_{\text{politeness}}$ model refutation, factuality and politeness respectively. $f_{\text{den}}(x, y)$ is used to evaluate claim-response relevance, and $\max\{f_{\text{den}}(e_i, y)\}_{i=1}^m$ evaluates the bestmatching evidence-response relevance. α is a hyperparameter to scale the relevance reward.

Reinforcement Learning. For reinforcement learning with human feedback, we use Equation (2) to assess the quality of the generated responses. The objective is to maximize the expected reward w.r.t. the actor model under the constraint that the parameters in π_{act} do not significantly deviate from π_{ref}. Formally, the learning is formulated as:

$$\max_{\pi_{\text{act}}} \mathbb{E}_{(x,e)\sim\{x_i,e_i\}_{i=1}^{n}, y\sim\pi_{\text{act}}(x,e)}[\hat{r}(x,e,y)] - \beta\mathbb{D}[\log\pi_{\text{act}}(y|x,e)||\log\pi_{\text{ref}}(y|x,e)]], \tag{3}$$

where β is a hyperparameter to regularize the output difference between π_{act} and π_{ref}, and $\mathbb{D}$ stands for the KL divergence. Here, we aim to find the optimal policy π_{act} that maximizes the expected reward $\hat{r}$. The additional KL divergence term controls how far the actor model can travel from the reference model via the minimization of KL divergence. Therefore, penalizing the KL distance effectively prevents optimization instability or model collapse. During training, we initialize the actor model π_{act} with reward head using the weights from π_{ref} and leverage PPO as the learning algorithm (Schulman et al., 2017; Ouyang et al., 2022). We provide an illustration of the training pipeline in Figure 3, where the upper subfigure demonstrates evidence retrieval and response generation. The lower subfigure explains our reinforcement learning scheme, in which the generated responses are used to estimate the rewards and compute the KL distances. Finally, the rewards and KL penalty are used the compute the training loss and update the actor model. After the reinforcement learning stage, the resulting actor model π_{act} is used as our response generation model in RARG.

4 Experiments

4.1 Experiment Design

Evidence Retrieval. Our retrieval module consists of BM25 and f_{den} (initialized with E5 (Wang et al., 2022)). We adopt the Check-COVID dataset for evidence retrieval evaluation (Wang et al., 2023a), the adopted metrics are NDCG and Recall (i.e., N@k and R@k) with $k \in [1, 3, 5]$. For baselines, we adopt sparse algorithms TFIDF and BM25 (Robertson et al., 1995). We also incorporate state-of-the-art dense retrievers DPR and E5 for comparison (Karpukhin et al., 2020; Wang et al., 2022).

Response Generation. We adopt our collected scientific articles (see Section 3.2) and use the top-3 / top-5 documents from retrieval. Our f_{gen} is based on Llama 2 (7B) (Touvron et al., 2023), with the supervised fine-tuning variant denoted by RARG (S). We adopt MisinfoCorrect dataset for training (He et al., 2023) and perform both in-domain and cross-domain evaluation (on Constraint and ANTiVax) (Patwa et al., 2021; Hayawi et al., 2022). We follow the evaluation from (He et al., 2023) using metrics refutation (R.), factuality (F.) and politeness (P.), we also evaluate the response relevance to input claim (C.) and evidence (E.) using our dense retriever (i.e., f_{den}). Baseline methods include BART, DialoGPT, PARTNER, GODEL, MisinfoCorrect (MisinfoC.), Llama 2 w/ and w/o retrieval (denoted with Llama and Llama (R)) and GPT-3.5 (Lewis et al., 2020a; Zhang et al., 2020; Sharma et al.; Peng et al., 2022; He et al., 2023; Touvron et al., 2023; Brown et al., 2020).

4.2 Evidence Retrieval Results

Our main retrieval results are reported in Table 1. In this table, rows represent retrieval methods and the columns represent different metrics. For top-1 scores, we use N@1 since top-1 NDCG and Recall scores are equivalent. From the results we observe: (1) RARG retriever consistently outperforms baseline retrieval methods across all metrics, with an average performance improvement of 7.09% com-

	Check-COVID Dataset				
	N@1 ↑	N@3 ↑	R@3 ↑	N@5 ↑	R@5 ↑
TFIDF	0.266	0.363	0.427	0.385	0.480
BM25	0.292	0.395	0.467	0.426	0.545
DPR	0.324	0.411	0.477	0.457	0.588
E5	<u>0.445</u>	<u>0.584</u>	<u>0.679</u>	<u>0.609</u>	<u>0.741</u>
RARG	**0.513**	**0.631**	**0.712**	**0.646**	**0.750**

Table 1: Evidence retrieval results, with best results in bold and second best results underlined.

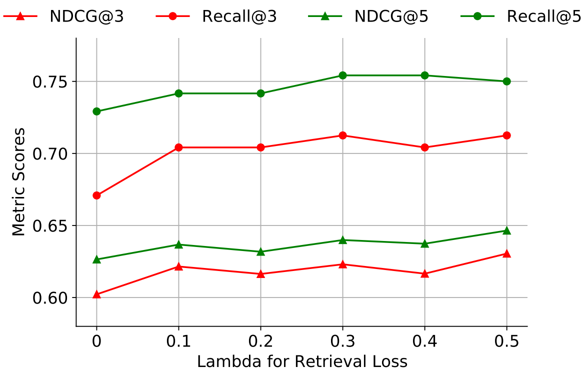

Figure 4: Hyperparameter sensitivity of λ in retrieval.

	MisinfoCorrect Dataset				
	R. ↑	F. ↑	P. ↑	C. ↑	E. ↑
BART	0.824	0.623	0.824	0.799	0.685
DialoGPT	0.831	0.693	0.874	0.800	0.689
PARTNER	0.792	0.779	0.790	0.759	0.663
GODEL	0.931	0.904	0.987	0.751	0.680
MisinfoC.	0.916	0.914	0.927	0.788	0.686
Llama	0.727	0.950	0.984	0.761	0.674
Llama (R)	0.761	0.942	0.984	0.776	0.694
GPT-3.5	0.857	**0.971**	0.987	0.805	0.690
RARG (S)	0.922	0.944	0.988	0.804	0.702
RARG	**0.965**	0.967	**0.989**	**0.812**	**0.704**

Table 2: In-domain response generation results, with best results in bold and second best results underlined.

pared to the second best results. (2) In contrast to sparse retriever along, the additional dense retriever significantly improves the ranking performance. For example, RARG achieves 37.61% performance improvement in Recall@5 compared to BM25. (3) The performance gains through our dense retriever increases as we narrow the size of the retried subset (i.e., top-k). For instance, the NDCG improvements compared to the best baseline rise from 6.07% to 8.04% with k decreasing from 5 to 3. (4) By leveraging the ranking margin and contrastive learning, our dense retriever successfully exploits noisy BM25 results and outperforms the second best baseline E5 by 15.28% in top-1 score, indicating strong reranking performance with RARG. Overall, we find the two-stage retrieval pipeline in RARG performs well even with limited training data, which demonstrates substantially improved retrieval performance.

We additionally study the effect of λ in the training objective of the retriever model, which regularizes the strength of contrastive loss. In particular, we vary the value of λ from 0 to 0.5 and evaluate the retrieval performance on NDCG and Recall with $k \in [3, 5]$. We focus on top-3 and top-5 scores as they are representative ranking scores and show consistent trends with changing λ. We present the performance visually in Figure 4, with x-axis representing the λ values and y-axis representing the ranking scores. We observe consistent improvements by applying the contrastive loss, followed by minor changes with futher increasing λ values. The performance for all metrics remains robust even with the maximum value of 0.5. In sum, the proposed contrastive loss improves the performance of RARG retrieval, and performs quite robust regardless of hyperparameter selections.

4.3 Response Generation Results

For response generation, we leverage the retrieved documents as part of the input prompt to generate the counter-responses. We report the in-domain response generation results in Table 2, with rows representing response generation models and columns representing different metrics. To evaluate the model generalization, we additionally perform response generation on cross-domain datasets Constraint and ANTiVax (Patwa et al., 2021; Hayawi et al., 2022), with results presented in Table 3. We also demonstrate the generation quality on cross-domain data via qualitative examples in Table 6.

From both in- and cross-domain evaluation results we observe: (1) Both RARG versions perform well on in-domain data and achieve superior performance over baseline methods. For example, RARG outperforms the best baseline on the refutation metric with 3.65% relative improvement. (2) RLHF-tuned RARG can further improve generation quality, achieving an average 2.40% performance improvement on refutation, factuality and politeness metrics compared to RARG (S). (3) Despite significantly reduced size (i.e., 7B), RARG performs similarly or exceeds GPT-3.5 on all metrics. In particular, RARG significantly outperforms GPT-3.5 on refutation, which may be attributed to the misinterpretation of certain examples by GPT-3.5, as we demonstrate in Table 6. (4) In cross-domain experiments (Table 3), although most models show comparable performance on refutation, factuality and politeness metrics, RARG can generate responses of substantially enhanced coherence and quality. In contrast, baseline methods exhibit signs of overfitting, leading to responses that highly resemble the training data (see Table 6). (5) Llama 2 and GPT-3.5 perform well in cross-domain response generation (Table 3). Yet they show significantly reduced refutation scores, likely stemming from the frequent misclassification (i.e., as if responding to valid information). (6) Combined both quantitative and qualitative results on cross-domain data, RARG clearly outperforms baseline generation methods with significantly improved claim and evidence relevance as well as generation quality. Overall, we conclude that the RLHF-tuned RARG demonstrates enhanced claim understanding, evidence-based reasoning and response generation abilities. Furthermore, RARG exhibits superior performance in countering both in-domain and cross-domain mis-

	Constraint Dataset					ANTiVax Dataset				
	R. ↑	F. ↑	P. ↑	C. ↑	E. ↑	R. ↑	F. ↑	P. ↑	C. ↑	E. ↑
BART	0.949	**0.969**	**0.987**	0.692	0.608	0.942	0.949	0.960	0.741	0.641
DialoGPT	0.953	0.953	0.968	0.636	0.584	0.952	0.952	0.966	0.748	0.641
PARTNER	0.952	0.953	0.968	0.633	0.585	0.887	0.882	0.901	0.732	0.632
GODEL	0.938	0.908	0.986	0.594	0.578	0.931	0.906	0.985	0.706	0.668
MisinfoC.	**0.955**	0.955	0.970	0.636	0.584	0.952	0.952	0.967	0.747	0.641
Llama	0.595	0.929	0.977	0.745	0.628	0.655	0.942	0.984	0.753	0.663
Llama (R)	0.708	0.930	0.976	0.750	0.634	0.722	0.943	0.984	0.775	0.689
GPT-3.5	0.855	0.933	0.979	0.795	**0.684**	0.909	0.948	0.987	0.780	0.689
RARG (S)	0.941	0.937	0.981	0.799	0.679	0.966	0.975	0.989	0.791	0.687
RARG	0.951	0.940	0.982	**0.805**	0.682	**0.971**	**0.978**	**0.991**	**0.795**	**0.691**

Table 3: Cross-domain response generation results, with best results in bold and second best results underlined.

information, highlighting the potential of RARG in generating counter-misinformation responses across a wide range of real-world scenarios.

5 Conclusion

In this paper, we propose a novel evidence-driven retrieval augmented response generation framework RARG against online misinformation. To the best of our knowledge, RARG is the first to introduce evidence-backed response generation to counter misinformation. The proposed RARG comprises of: (1) evidence retrieval, where evidence documents are efficiently collected and reranked; and (2) response generation, in which RARG generates evidence-based counter-responses that are factual and polite. We demonstrate the effectiveness of RARG by performing extensive experiments on multiple datasets, where RARG can consistently generate evidence-based responses with improved quality over state-of-the-art baseline methods.

6 Limitations

Despite introducing RARG for evidence-based counter-response generation against online misinformation, we have not discussed the setting where test domains significantly differ from COVID (e.g., fake news), which may hinder the deployment of the proposed method for more generalized applications. In addition, we have not studied the case when the reliability of evidence sources can not be guaranteed (e.g., documents from unverifiable online sources), which may introduce inaccuracies and impact the validity of the generated responses. Due to the limited research scope and budgets, we have not conducted human evaluations to assess the overall response quality of RARG. Therefore, we plan to explore a more generalized and domain-adaptive solution with an improved evaluation protocol for counter-response generation in the future.

Acknowledgment

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105032, IIS-2130263, CNS-2131622, CNS-2140999. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Philip Ball and Amy Maxmen. 2020. The epic battle against coronavirus misinformation and conspiracy theories. *Nature*, 581(7809):371–375.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of two crowds: Misinformation moderation on reddit and how to improve this process—a case study of covid-19. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–33.

Erik Brand, Kevin Roitero, Michael Soprano, and Gianluca Demartini. 2021. E-bart: Jointly predicting and explaining truthfulness. In *TTO*, pages 18–27.

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

Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11):1531–1546.

Canyu Chen and Kai Shu. 2023a. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Canyu Chen and Kai Shu. 2023b. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Canyu Chen, Haoran Wang, Matthew Shapiro, Yunyu Xiao, Fei Wang, and Kai Shu. 2022. Combating health misinformation in social media: Characterization, detection, intervention, and open issues. *arXiv preprint arXiv:2211.05289*.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2021. Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Chiara Patricia Drolsbach and Nicolas Pröllochs. 2023. Diffusion of community fact-checked misinformation on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Kadhim Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public health*, 203:23–30.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021b. Compare to the knowledge: Graph neural fake news detection with external knowledge. 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 754–763, Online. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Gongyao Jiang, Shuang Liu, Yu Zhao, Yueheng Sun, and Meishan Zhang. 2022. Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, 59(5):103029.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Boshko Koloski, Timen Stepišnik Perdih, Marko Robnik-Šikonja, Senja Pollak, and Blaž Škrlj. 2022. Knowledge graph informed fake news classification via heterogeneous representation ensembles. *Neurocomputing*.

Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022a. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–25.

Ziyi Kou, Lanyu Shang, Yang Zhang, Christina Youn, and Dong Wang. 2021. Fakesens: A social sensing approach to covid-19 misinformation detection on social media. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 140–147. IEEE.

Ziyi Kou, Lanyu Shang, Yang Zhang, Zhenrui Yue, Huimin Zeng, and Dong Wang. 2022b. Crowd, expert & ai: A human-ai interactive approach towards natural language explanation based covid-19 misinformation detection. In *Proc. Int. Joint Conf. Artif. Intell.(IJCAI)*, pages 5087–5093.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2017. Efficient and timely misinformation blocking under varying cost constraints. *Online Social Networks and Media*, 2:19–31.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, Toronto, Canada. Association for Computational Linguistics.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Pranav Malhotra, Kristina Scharp, and Lindsey Thomas. 2022. The meaning of misinformation and those who correct it: An extension of relational dialectics theory. *Journal of Social and Personal Relationships*, 39(5):1256–1276.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835, Toronto, Canada. Association for Computational Linguistics.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. In *2020 IEEE international Conference on big data (big data)*, pages 748–757. IEEE.

Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 651–662.

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29. Springer.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. *Scientific reports*, 12(1):5966.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183, Online. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Nikita Mariam Santhosh, Jo Cheriyan, and Lekshmi S Nair. 2022. A multi-model intelligent approach for rumor detection in social networks. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–5. IEEE.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. 2022. If you have a reliable source, say something: effects of correction comments on covid-19 misinformation. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 896–907.

Lanyu Shang, Ziyi Kou, Yang Zhang, Jin Chen, and Dong Wang. 2022a. A privacy-aware distributed knowledge graph approach to qois-driven covid-19 misinformation detection. In *2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022b. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference 2022*, pages 3623–3631.

Lanyu Shang, Yang Zhang, Zhenrui Yue, YeonJung Choi, Huimin Zeng, and Dong Wang. 2022c. A knowledge-driven domain adaptive approach to early misinformation detection in an emergent health domain on social media. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 34–41. IEEE.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach.

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

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in Fake Media Generation and Detection*, pages 215–232. Springer.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. *Disinformation, misinformation, and fake news in social media*. Springer.

Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 proceedings*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Bruno Tafur and Advait Sarkar. 2023. User perceptions of automatic fake news detection: Can algorithms fight online misinformation? *arXiv preprint arXiv:2304.07926*.

Yuko Tanaka and Rumi Hirayama. 2019. Exposure to countering messages online: alleviating or strengthening false belief? *Cyberpsychology, Behavior, and Social Networking*, 22(11):742–746.

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

Jeyasushma Veeriah. 2021. Young adults' ability to detect fake news and their new media literacy level in the wake of the covid-19 pandemic. *Journal of Content, Community and Communication*, 13(7):372–383.

Nguyen Vo and Kyumin Lee. 2019. Learning from fact-checkers: Analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344.

Nguyen Vo and Kyumin Lee. 2020. Standing on the shoulders of guardians: Novel methodologies to combat fake news. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 183–210.

Emily K Vraga and Leticia Bode. 2021. Addressing covid-19 misinformation on social media preemptively and responsively. *Emerging infectious diseases*, 27(2):396.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023a. Check-COVID: Fact-checking COVID-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14114–14127, Toronto, Canada. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv e-prints*, pages arXiv–2212.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023b. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Junfei Wu, Qiang Liu, Weizhi Xu, and Shu Wu. 2022a. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

Junfei Wu, Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2022b. Adversarial contrastive learning for evidence-aware fake news detection with graph neural networks. *arXiv preprint arXiv:2210.05498*.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022c. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 2501–2510.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. MetaAdapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5223–5239, Toronto, Canada. Association for Computational Linguistics.

Huimin Zeng, Zhenrui Yue, Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022. Unsupervised domain adaptation for covid-19 information service with contrastive adversarial domain mixup. pages 159–162.

Huimin Zeng, Zhenrui Yue, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Unsupervised domain adaptation via contrastive adversarial domain mixup: A case study on covid-19. *IEEE Transactions on Emerging Topics in Computing*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Runcong Zhao, Miguel Arana-catania, Lixing Zhu, Elena Kochkina, Lin Gui, Arkaitz Zubiaga, Rob Procter, Maria Liakata, and Yulan He. 2023. PANACEA: An automated misinformation detection system on COVID-19. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE.

Jianming Zhu, Smita Ghosh, and Weili Wu. 2021. Robust rumor blocking problem with uncertain rumor sources in social networks. *World wide web*, 24:229–247.

Datasets	#Evid.	#Claim	#Resp.	#Leng.
Check-COVID	5.6k	1,068/229	N/A	19.4
MisinfoC.	N/A	568/134	568	38.8
ANTiVax	N/A	500	N/A	37.2
Constraint	N/A	500	N/A	25.0

Table 4: Dataset statistics.

	BA ↑	Acc. ↑	F1 ↑	Prec. ↑	Rec. ↑
Refutation	0.835	0.800	0.873	0.978	0.788
Factuality	0.928	0.933	0.942	0.925	0.961
Politeness	0.849	0.881	0.923	0.941	0.905

Table 5: Reward modeling results.

A Experiment Details

A.1 Evidence Retrieval

Our two-stage retrieval consists of BM25 and a fine-tuned dense retriever. The dense retriever is initialized with the pretrained embeddings from bidirectional encoder representations (E5) (Wang et al., 2022). We describe the experimental design below.

- *Dataset*: For retrieval evaluation, we adopt the Check-COVID dataset that provides claim-evidence pairs (Wang et al., 2023a). The dataset statistics of the training/test set is provided in Table 4.

- *Baseline*: For baseline models, we adopt unsupervised sparse retrieval algorithms TFIDF and BM25 (Robertson et al., 1995). We also incorporate two state-of-the-art dense retrievers for comparison: dense passage retrieval (DPR) and E5 retriever (Karpukhin et al., 2020; Wang et al., 2022).

- *Evaluation*: In our evaluation, we adopt the following evaluation ranking metrics: normalized discounted cumulative gain (NDCG@k) and recall (Recall@k) with $k \in [1, 3, 5]$. We validate the model using the best NDCG@10 scores and rerank the retrieved results.

For the subsequent response generation experiments, we adopt our collected scientific articles as the evidence sources (see Section 3.2) and collect the top-3 / top-5 documents for each input claim.

A.2 Response Generation

For experiments in response generation, the base model is the 7B version of Llama 2 (Touvron et al., 2023). To improve parameter efficiency, RARG is trained with 8bit quantization and LoRA (Hu et al., 2021a; Dettmers et al., 2022), we report the experiment design below.

- *Dataset*: We adopt MisinfoCorrect for training (He et al., 2023). All models are trained with the identical training set and evaluated on a non-overlapping test set for unbiased assessment. To evaluate model generalization, we additionally adopt Constraint and ANTiVax, with 500 sampled examples each, see Table 4 (Patwa et al., 2021; Hayawi et al., 2022).

- *Baseline*: For baseline models, we adopt multiple state-of-the-art baselines. In particular, we adopt text generation-based BART, DialoGPT, GODEL (Lewis et al., 2020a; Zhang et al., 2020; Peng et al., 2022), reinforcement learning-based methods PARTNER and MisinfoCorrect (Sharma et al.; He et al., 2023), LLM-based Llama 2, Llama 2 with retrieval (with identical retrieval as in RARG, denoted with Llama (R)) and GPT-3.5 (Brown et al., 2020).

- *Evaluation*: We follow the evaluation protocol from, with metrics in refutation, factuality and politeness using the supervised trained LMs in Section 3.3 (He et al., 2023). We also evaluate the response relevance to input claims and the collected evidence using the relevance estimation model (i.e., dense retriever) as in Section 3.2.

Besides RARG, we also report results of the supervised fine-tuned version of RARG (i.e., RARG (S)). For reward modeling, we train three BERT models on the binary human feedback data w.r.t refutation, factuality and politeness. As the label distribution is imbalanced, we adopt stratified sampling to split train/test sets and apply class-balanced cross entropy loss for training. Besides accuracy, F1, precision and recall, we also use balanced accuracy (BA) to evaluate class-balanced performance on human feedback classification. The evaluation results for reward modeling are in Table 5. In summary, BERT models achieve well-balanced classification performance across all human feedback metrics, with average BA of 0.871 and accuracy of 0.871.

A.3 Implementation

For retrieval baseline methods, we follow the original works for implementation and hyperparameter selection. For RARG retrieval pipeline, we use the E5-base version for further tuning and train with AdamW optimizer for 5 epochs. We search the learning rate from [1e-5, 2e-5, 3e-5] and select both τ and λ from $[0.1, 0.2, 0.3, 0.4, 0.5]$. Training is performed with 100 warm up steps and cosine learning rate scheduler. For sampling evidence, 4 positive evidence documents and 4 negative documents are adopt for training. We align the rest settings with the original E5 work (Wang et al., 2022). For the generation pipeline, we adopt publicly available implementation from the original authors. To train our 7B version Llama 2, we adopt 8bit quantization and LoRA for parameter efficient tuning, which is less than 0.2% of the original parameters in Llama 2 (Hu et al., 2021a; Dettmers et al., 2022). We use consistent LoRa settings with 16 as LoRA dimension, 32 as LoRA α along with 0.05 for LoRA dropout. The LoRA target modules are the Q and V projection matrices. For supervised fine-tuning, we train 3 epochs with learning rate of 1e-4 and 0.03 percentage of the steps as warmup. For reinforcement tuning, we adopt PPO with 1e-5 learning rate and 0.2 as initial KL regularization. Training epochs is selected between 1 to 3, batch size is selected between 4 to 32, and we update parameters after 4 gradient accumulation steps. To construct the text input for instruction tuning, we prepend evidence as context, followed by an instruction that describes the task. We also incorporate the input misinformation in the prompt before the start of the response token. The resulting prompt template is:

```
### Instruction
{evidence}; Based on the above evidence, determine if the claim is valid and
explain why: {claim}
### Response
{response}
```

where `evidence` and `claim` are filled by the collected evidence and claim. For training, the annotated response of each data example follows. For inference, the `response` position is left empty for generation.

B Qualitative Results

We provide additional qualitative examples to compare the response quality among different response generation methods, examples can be found in Table 6. Due to space limitations, we select GODEL, MisinfoCorrect, Llama 2 (R) and GPT-3.5 for comparison, as the generated responses from such models are of the highest quality. Since RARG (S) and RARG generate similar responses to most claims, we provide qualitative example with RLHF-tuned version of RARG. Overall, we notice high similarity in the responses from GODEL and MisinfoCorrect compared to the training data. For LLMs, we observe significantly increased response length by Llama 2 (R) and GPT-3.5, which may affect the non-redundancy and readability scores in the response quality evaluation. In addition, Llama 2 (R) and GPT-3.5 often demonstrate false responses (i.e., hallucination) by taking the input claim as valid. For example in the example 1, both Llama 2 and GPT-3.5 incorrectly identifies the claim as valid, and thus generates invalid responses to the original misinformation. Similarly in example 2, GPT-3.5 produces questionable rationales on the efficacy of masks in combating COVID, possibly due to the lack of claim understanding. Considering both the model size and efficiency, RARG shows improved response generation capabilities and can outperform baseline methods w.r.t. the response quality.

	Response Examples from Constraint
Claim	*At a time when studies are linking air pollution with COVID-19 mortality rates for the most vulnerable in society, the Trump Administration is rolling back environmental protections - allowing more poison into air.*
GODEL	The statement that the vaccine manipulates the DNA is incorrect. As verified by fact-checkers, the vaccine ...
MisinfoC.	This is not correct. The vaccine is not related to COVID-19. The vaccine only uses mRNA for ...
Llama (R)	The claim made in the statement is **valid**. There have been several studies conducted that suggest a ...
GPT-3.5	The claim is **valid**. Studies have shown a correlation between air pollution and COVID-19 mortality rates ...
RARG	This is not correct. The air pollution has nothing to do with the mortality rate. The mortality rate is related to the virus. The air pollution is not the cause of the virus. Please stop this misinformation.
Claim	*Says a warning label on a box of disposable masks shows that they are ineffective at protecting against the spread of COVID-19.*
GODEL	The assertion is wrong. The vaccine is not the gene therapy and the nurse is not committing crime. The ...
MisinfoC.	The vaccine is not a vaccine. It is safe, and it is well tested on many people. Please stop spreading ...
Llama (R)	The claim that a warning label on a box of disposable masks shows that they are ineffective at protecting ...
GPT-3.5	It is **impossible to determine the validity** of this claim without seeing the actual warning label on the ...
RARG	The warning label is not correct. The masks are effective at protecting against the spread of COVID-19.
	Response Examples from ANTiVax
Claim	*People who don't care about Covid: You can't live in fear of the virus. Also people who don't care about Covid: I'm not getting a vaccine. It's experimental. We don't know what's in it. Me: So about living in fear?*
GODEL	Your understanding about the experiment is inaccurate. The vaccine is not the gene therapy, and it does ...
MisinfoC.	This is not correct. The vaccine is not experimental, and it is not the experimental gene therapy. It only ...
Llama (R)	The claim that people who don't care about COVID-19 are living in fear of the virus is not valid. It is ...
GPT-3.5	The claim is not valid. The statement highlights a contradiction in the attitudes of people who claim they ...
RARG	This is not correct. The vaccine is not experimental. It is well tested and safe.
Claim	*Firstly, how long have you worked there? 2nd If you have been there more than 3 years you will have more rights. 3rd If you don't want this untested vaccine, find examples of people who have had heart attacks and died because of vaccine. 4th Mention your human rights*
GODEL	Your understanding is not correct. The vaccine is safe, and before we take it, it has been tested on many ...
MisinfoC.	Your claim is not correct. The vaccine is not untested and has not been found to cause heart problems ...
Llama (R)	The claim you provided is not valid and is based on a series of false assumptions and misleading statement ...
GPT-3.5	This claim is not valid. It contains several unrelated statements and does not present a clear argument ...
RARG	Please stop spreading misinformation. The vaccine is not untested. It has been tested on thousands of people. It is safe and effective.

Table 6: Generated responses for baseline methods and RARG on Constraint and ANTiVax datasets.