F^{2}RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation

Warning: this paper discusses and contains content that can be offensive or upsetting.

Haiyang Wang[1*], **Yuchen Pan**[2*], **Xin Song**[1], **Xuechen Zhao**[1], **Minghao Hu**[3†], **Bin Zhou**[1†]
[1]National University of Defense Technology, Changsha, China
[2] Intelligent Game and Decision Lab, Academy of Military Science, Beijing, China
[3] Center of Information Research, Academy of Military Sciences, Beijing, China
{wanghaiyang19, panyuchen, songxin, zhaoxuechen, binzhou}@nudt.edu.cn
huminghao16@gmail.com

Abstract

Hate speech (HS) on social media exacerbates misinformation and baseless prejudices. Evidence-supported counterspeech (CS) is crucial for correcting misinformation and reducing prejudices through facts. Existing methods for generating evidence-supported CS often lack clear guidance with a core claim for organizing evidence and do not adequately address factuality and faithfulness hallucinations in CS within anti-hate contexts. In this paper, to mitigate the aforementioned, we propose F^{2}RL, a **F**actuality and **F**aithfulness **R**einforcement **L**earning framework for generating claim-guided and evidence-supported CS. Firstly, we generate counter-claims based on hate speech and design a self-evaluation mechanism to select the most appropriate one. Secondly, we propose a coarse-to-fine evidence retrieval method. This method initially generates broad queries to ensure the diversity of evidence, followed by carefully reranking the retrieved evidence to ensure its relevance to the claim. Finally, we design a reinforcement learning method with a triplet-based factuality reward model and a multi-aspect faithfulness reward model. The method rewards the generator to encourage greater factuality, more accurate refutation of HS, consistency with the claim, and better utilization of evidence. Extensive experiments on three benchmark datasets demonstrate that the proposed framework achieves excellent performance in CS generation, with strong factuality and faithfulness.

1 Introduction

Hate speech (HS) is an aggressive expression that incites hatred towards specific groups based on their group identity (religion, ethnicity, race, etc.) (Nockleby, 2000). The spread of HS on social media fuels misinformation and baseless prejudices (Waldron, 2012; Dreißigacker et al., 2024).

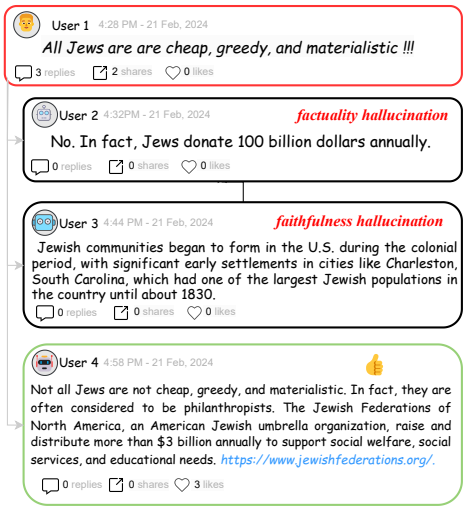

Figure 1: Examples of Different Evidence-supported Counterspeech.

Counterspeech (CS) involves directly responding to HS to reduce its negative impact and promote a more friendly and harmonious dialogue (Chung et al., 2023). The types of CS are diverse, including humor, rhetorical questions, evidence-supported, and others (Gupta et al., 2023; Wang et al., 2024). In particular, evidence-supported counterspeech is crucial for correcting misinformation, reducing baseless prejudices, and educating the audience through facts as evidence (Benesch et al., 2016).

The automatic generation of evidence-supported CS has been extensively researched. These studies can be categorized into **non-retrieval-augmented methods** and **retrieval-augmented methods**. Non-retrieval-augmented methods typically generate evidence-supported CS that relies on the internal parameterized knowledge of large language models (LLMs). He et al. (2023) proposes a reinforcement learning-based framework called MisinfoCorrect, which employs a BERT-based classifier as a reward model to enhance the factuality of the counter-responses. Wang et al. (2024) applies a discrimina-

[*]Equal contribution.
[†]Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4457–4470
November 12-16, 2024

tor to guide the decoding process of LLMs. Gupta et al. (2023) uses category distribution learning for LLMs to generate evidence-supported CS, as opposed to other types of CS. As for the retrieval-augmented methods, Chung et al. (2021) present a knowledge-grounded CS generation pipeline that uses an external knowledge base. The method constructs data pairs of HS and background knowledge. This allows for supervised fine-tuning of LLMs to generate evidence-supported CS. This method allows CS to include more up-to-date and factually correct knowledge. Recently, Yue et al. (2024) proposes a retrieval-augmented response generation (RARG) for online misinformation. RARG collects and reranks evidence from a large academic database, then uses PPO-based reinforcement learning to fine-tune LLMs for generating evidence-supported responses.

The aforementioned studies have advanced the development of evidence-supported CS generation, but they may still have the following limitations. **(L1) Factuality hallucination and evaluation challenges**: non-retrieval-augmented methods typically rely on the internal knowledge of LLMs. Therefore, the generated CS often contains factual errors (e.g., the CS by user 2 in Figure 1). Additionally, current methods use classifiers to evaluate the factuality of CS (He et al., 2023; Yue et al., 2024), which often lack objectivity and generalizability. **(L2) Faithfulness hallucination in CS**: existing research defines faithfulness hallucination as being inconsistent with the input content (Hu et al., 2024). In evidence-supported CS generation, this primarily manifests in two aspects: the inability to effectively rebut the HS and the failure to correctly utilize the given evidence (e.g., the CS by user 3 in Figure 1). **(L3) Evidence lacks the guidance of a clear claim**: existing methods tend to simply list evidence but lack the guidance of a clear claim. This may result in CS lacking a coherent argument and clear evidence connection. The CS from user 4 in Figure 1 is a good example. It presents a clear claim, followed by supporting evidence.

In this paper, to mitigate the aforementioned limitations, we propose **F^{2}RL**, a **F**actuality and **F**aithfulness **R**einforcement **L**earning framework to generate claim-guide evidence-supported CS. The framework first generates a counter-claim based on the HS, which serves as the core argument of the CS. Then, several queries are generated based on this claim to retrieve supporting evidence. Finally, given the claim and evidence, we optimize the generator using reinforcement learning to enhance the factuality and faithfulness of the CS.

Particularly, our model consists of three modules: **(1) Self-evaluation claim generation:** This module employs an LLM-based claim generator to produce various claims. The LLM then self-evaluates these claims to select the most appropriate one. **(2) Coarse-to-fine evidence retrieval:** This module generates queries based on the selected claim and uses a coarse-to-fine retrieval strategy to obtain supporting evidence for the claim. **(3) Factuality and faithfulness reinforcement learning :** This module trains a CS generator to generate claim-guided and evidence-supported CS. Specifically, We design a triplet-based factuality reward model and a multi-aspect faithfulness reward model to evaluate the generated CS. Then we use reinforcement learning to optimize the generator to improve the factuality and faithfulness. Experiments demonstrate that our framework outperforms strong baselines in the evidence-supported counterspeech generation task. Our contributions are threefold:

- We design a novel claim-guided coarse-to-fine evidence retrieval method. It first generates broad queries to ensure the diversity of evidence, then carefully rerank the results to ensure their relevance to the claim. This method enhances the coherence and evidence connection of the CS by closely aligning the evidence with the central counter-claim.

- We propose an innovative factuality and faithfulness reinforcement learning framework for claim-guided evidence-supported CS generation. It enables generating CS with higher factual correctness, more precise refutation, and better utilization of evidence, leveraging a triplet-based factuality reward model and a multi-aspect faithfulness reward model.

- Extensive experiments on 3 benchmark datasets show that the proposed framework achieves excellent performance in CS generation with good factuality and faithfulness. It also generalizes well to different LLMs.

2 Related Work

2.1 Counterspeech Generation

Counterspeech (**CS**) can be defined as a direct response to hate or dangerous speech to mitigate hate. CS can fight hate speech (**HS**) and reduce

its negative impact on social media while still allowing free speech (Chung et al., 2023). Recently, many automatic counterspeech generation methods have been proposed. Zhu and Bhat (2021) propose Generate-Prune-Select which is a three-stage pipeline to obtain the most relevant CS for an HS instance. Chung et al. (2021) proposed a knowledge-grounded generation approach by incorporating an intermediate step in which keyphrases are generated to retrieve the necessary knowledge. Saha et al. (2022) proposed CounterGEDI, an ensemble of GEDI to guide the generation of a DialoGPT model toward more polite, detoxified, and emotional CS. Then, Gupta et al. (2023) proposed QUARC, which leverages vector-quantized representations to generate CS with various intent categories. Jiang et al. (2023) proposed RAUCG, which enhances the LM's ability to automatically incorporate counter-knowledge from new external statistics, facts, or examples in counter-narrative generation. Wang et al. (2024) proposed DART, which employed dual discriminator to jointly guide the decoding preferences of LLMs, aiming to generate CS catering to specific intent and hate mitigation. However, these methods focus on improving CS quality and diversity, often overlooking the importance of ensuring the factuality and faithfulness of CS.

2.2 Reinforcement Learning for LLMs

LLMs acquire surprising capabilities (Touvron et al., 2023), largely due to the fine-tuning of LLMs using Reinforcement Learning from Human Feedback (RLHF). Recently, RLHF has become key in fine-tuning LLMs to better align with human preferences and improve task performance (Christiano et al., 2017). RLHF generally includes four processes (Lang et al., 2024): supervised fine-tuning, human preference collecting, reward learning and RL policy optimization. Currently, two main RLHF approaches are reward-based methods and reward-free methods. OpenAI pioneered the reward-based approach, utilizing preference data to construct a reward model and optimizing the reward signal with actor-critic algorithms like Proximal Policy Optimization (PPO) (Schulman et al., 2017). Conversely, reward-free methods dispense with the explicit use of a reward function. For example, DPO (Rafailov et al., 2023) represents the reward function in the logarithmic form of the policy and focuses solely on policy optimization. Other reward-free methods include RRHF (Yuan et al., 2023) and PRO (Song et al., 2024).

3 Methodology

3.1 Task Definition and Pipeline

Formally, the goal of the claim-guided evidence-supported CS generation task is to construct a stochastic text generation function χ. It can take hate speech x, claim c and reference evidence set E as the input and output the generated CS $\hat{y}$. The core argument of the CS $\hat{y}$ needs to align with the given claim and effectively utilize the provided evidence to counter HS such that $\hat{y} \sim \chi(\cdot|x, c, E)$.

3.2 Model Architecture

In this section, we describe the main modules of our proposed F^{2}RL framework for claim-guided evidence-supported counterspeech generation. As demonstrated in Figure 2, the proposed F^{2}RL framework mainly consists of three modules:

- **Self-Evaluation Claim Generation Module** employs an LLM-based claim generator to generate various claims. Subsequently, we design a voting prompt for the LLM, enabling it to compare different claims and vote for the one that most effectively rebuts HS.
- **Coarse-to-Fine Evidence Retrieval Module** retrieves and reranks documents to get the relevant supporting evidence for the claim. We also optimize the ranker using contrastive learning to improve the relevance estimation between the claim and the evidence.
- **Factuality and Faithfulness Reinforcement Learning Module** aims to train a CS generator to generate claim-guided and evidence-supported CS. This module applies a triplet-based factuality reward model and a multi-aspect faithfulness reward model to estimate the rewards of CS and update the parameters of the generator using PPO-based reinforcement learning.

3.3 Self-Evaluation Claim Generation

This module primarily focuses on generating a counter-claim based on the HS. It can explicitly expose the error of the input HS. The module consists of two sub-modules: **(1) Claim Generation**, which leverages the in-context learning (ICL) ability of LLMs to generate various counter-claims; and **(2) Self-Evaluation**, which applies a voting prompt for the LLM to compare different partial

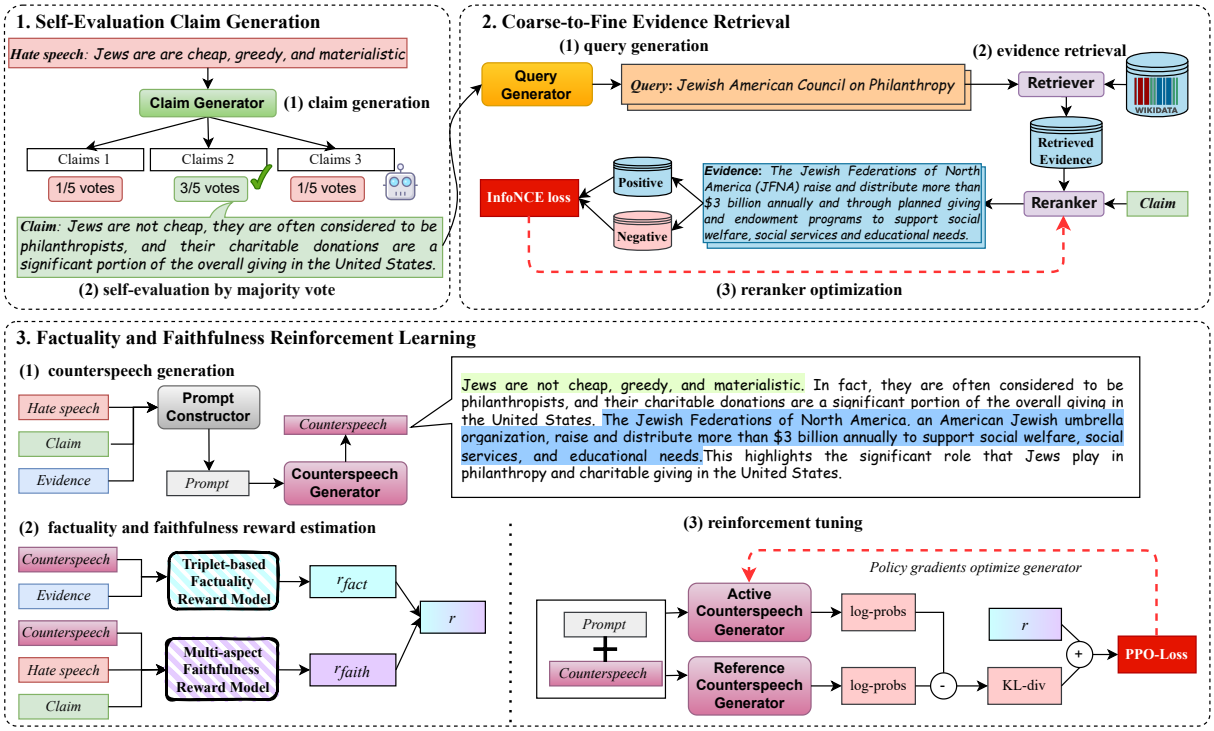

Figure 2: The architecture of the proposed **F^{2}RL** framework. (1) Self-Evaluation claim generation module generates and selects a counter-claim. (2) Coarse-to-Fine evidence retrieval module obtains the supporting evidence. (3) Factuality and faithfulness reinforcement learning module Optimize a CS generator.

claims and vote for the most promising one. Then, it employs a majority voting strategy to obtain more robust results (Yao et al., 2023).

Specifically, given the HS x, we design an instruction prompt template for an LLM to obtain a claim generator $\mathcal{G}_{claim}$. LLMs have the ICL capability (Brown et al., 2020), allowing them to perform claim generation without fine-tuning when provided with instructions and some examples. This generator takes the HS as input and outputs a series of claims $C = \{c_0, c_1, \cdots, c_{n_c}\}$, where n_c is the number of claims. Then, we design a voting prompt for the LLM to obtain a voting agent $\mathcal{V}$. This agent takes a series of claims C as input and selects the best claim c^*, which can be formulated as $c^* = \mathcal{V}(C)$. c^* is selected based on deliberately comparing different claims in C in the vote prompt. When using LLMs for self-evaluation, performing the process multiple times usually yields a more robust result. Therefore, we conduct multiple rounds of voting to select the claim with the most votes. Through self-evaluation, we hope to select claims that explicitly and objectively expose the errors or biases in HS more effectively than other claims, as the basis for subsequent evidence retrieval.

3.4 Coarse-to-Fine Evidence Retrieval

This module takes a query-based coarse-to-fine strategy to retrieve supporting evidence.

3.4.1 Query generation

Similar to claim generation, we design an instruction prompt for the LLM to obtain a query generator $\mathcal{G}_{query}$. This generator takes the claim c as input and generates a set of queries $Q = \{q_1, q_2, \cdots, q_{n_q}\}$, where n_q is the number of queries. This process can be formalized as $Q = \mathcal{G}_{query}(c)$. Generating queries based on claims offers several advantages over directly using claims or HS for retrieval (Zhao et al., 2024; Huang and Huang, 2024). It captures different aspects of the claim, thus increasing the relevance, comprehensiveness, and diversity of the evidence obtained.

3.4.2 Evidence Retrieval

This sub-module aims to retrieve evidence that can support the claim. Inspired by previous work (Yue et al., 2024), we designed a coarse-to-fine evidence retrieval pipeline. This pipeline first uses a retriever to conduct a broad initial retrieval based on queries, ensuring the diversity and comprehensiveness of the evidence set. Then, it employs a reranker to

perform fine-grained reranking, ensuring high relevance to the claim (Huang and Huang, 2024).

In particular, for the retriever $\mathcal{R}$, we use the off-the-shelf Contriever-MS MARCO (Izacard et al., 2022; Asai et al., 2023) by default to retrieve relevant documents from the Wikipedia database. Subsequently, we obtain an initial set of retrieved evidence, denoted as E_{coarse}. These processes can be formalized as $E_{\text{coarse}} = \mathcal{R}(Q, Wiki)$. The set E_{coarse} contains all evidence documents retrieved based on all queries Q. For the reranker process $\mathcal{R}_{rank}$, to obtain evidence that is relevant and supportive of the claim c, we employ the BGE M3 (Chen et al., 2024) model to calculate the relevance score. Based on this score, we rerank and filter E_{coarse} to get the fine-grained evidence set E_{fine}, which can be formalized as $E_{\text{fine}} = \mathcal{R}_{rank}(E_{\text{coarse}}, c)$.

3.4.3 Reranker Optimization

To improve the ranking performance and generalizability, We optimize the reranker's performance after each retrieval. It can be divided into two steps: (1) sampling of positive and negative evidence, and (2) optimization based on contrastive learning.

First, for each claim $c_i \in C$ and the initial evidence set E_{coarse}, we use the BGE model to calculate relevance scores. By setting different thresholds, we sample K positive evidence $\{e_j^p\}_{j=1}^k$ and K negative evidence $\{e_j^n\}_{j=1}^k$, forming positive pairs (c, e^p) and negative pairs (c, e^n). Subsequently, we minimize the InfoNCE loss (Chen et al., 2020; Yue et al., 2024) to draw the positive claim-evidence pairs closer and push away the negative samples. InfoNCE loss is a contrastive learning loss used to optimize representation learning by maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs. The optimization objective $\mathcal{L}$ can be formulated as:

$$\mathcal{L} = \sum_{c_i \in C} \sum_{e_j \in E_j^p} \ell_{(i,j)} \tag{1}$$

$$\ell_{(i,j)} = -\log \frac{\exp\left(s\left(c_i, e_j^p\right)/\tau\right)}{\sum_{e_k \in \left\{E_i^p, E_i^n\right\}} \exp(s(c_i, e_k)/\tau)} \tag{2}$$

where $s(\cdot,\cdot)$ is the BGE M3-based similarity function and τ denotes a temperature paramete.

3.5 Factuality and Faithfulness Reinforcement Learning

This section aims to use factuality and faithfulness reinforcement learning to fine-tune a CS generator.

3.5.1 Counterspeech Generation

This sub-module aims to generate CS using an LLM-based generator by designing prompt templates. Since different prompts may affect performance (Qian et al., 2023), we strive to maintain a consistent style when designing the prompts. The devised prompt template for CS generation consists of the following components:

> Task definition + Instruction + Hate speech x + Claim c+ Evidence E.

Among them, the task definition provides a standard definition of the CS generation task. The instruction specifies the guidelines we expect the generator to follow. The prompt can be formalized as $p = Prompt(x, c, E)$, where the $Prompt(\cdot)$ is the prompt template. Next, we use a supervised fine-tuned CS generator $\mathcal{G}_{cs}$ to generate CS cs based on the prompt p.

3.5.2 Factuality and Faithfulness Reward Estimation

Triplet-based Factuality Reward Model. The remarkable fluency and inventiveness of LLMs have made them popular (Zhao et al., 2023). Nonetheless, LLMs often generate persuasive but incorrect statements, known as hallucinations (Huang et al., 2023). Factuality hallucinations involve claims contradicted by real-world facts. Preventing factuality hallucinations in generating CS for HS is crucial, as these errors can undermine the CS's effectiveness. We design a triplet-based factuality reward model to address this limitation inspired by previous work (Hu et al., 2024).

Conditioned on the retrieved evidence as a premise, we define factuality reward as the probability that the CS is entailed by the retrieved evidence. Specifically, we first define a triple extractor $Etr(\cdot)$ that takes the generated CS cs_i as input and extracts knowledge triplets $T_i = \{(s_{ij}, r_{ij}, o_{ij})\}_{j=1}^{N_i}$, where N_i is the number of triplets. It can be formalized as $T_i = Etr(cs_i)$. Existing research has shown that (Wang et al., 2023; Hu et al., 2024) the decomposition of the original text into triplets facilitates finer-grained factuality

hallucination detection and more accurate factuality evaluation. Next, we construct evidence-triplet pairs (e, t), where $e \in E_i$ and $t \in T_i$. We employ an LM-based factuality checker, denoted as $Ckr(\cdot)$ to calculate the likelihood of entailment for each evidence-triplet pair. Finally, we use the average likelihood of entailment across all evidence-triplet pairs as the factuality reward. It can be formalized as follows:

$$r_{fact}(cs_i) = \frac{1}{N_E \cdot N_T} \Sigma_{e \in E_i} \Sigma_{t \in T_i} Ckr(e, t), \tag{3}$$

where N_E is the number of evidence and N_t is the number of triplets.

Multi-aspect Faithfulness Reward Model. Faithfulness means consistent with the input content (Li et al., 2022). As for the claim-guide evidence-supported CS generation, faithfulness has multiple aspects of meaning: **(1) Faithfulness to Hate Speech**: The CS must directly address and rebut the input HS, clearly pointing out its errors, biases, or inaccuracies. **(2) Faithfulness to the Claim**: The CS should be consistent with the input claim, revolving around it and ensuring it stays true to the claim's main points and facts. **(3) Faithfulness to the Evidence**: The CS must accurately reference and interpret the provided evidence to support the claim, enhancing its persuasiveness and credibility.

In detail, for the faithfulness to HS, we trained a binary stance detection model f_{hate} that can use HS as the target to evaluate the stance of a given text, identifying whether it supports or opposes the HS. We use the probability value of the opposing stance as the reward for faithfulness to HS. Then, we use a pre-built similarity function to measure the relevance of the CS to the claim and the evidence. We use the sum of the above scores as the faithfulness reward r_{faith}, which can be formulated as:

$$\begin{aligned} r_{faith}(cs_i) = & f_{hate}(x_i, cs_i) + s(c_i, cs_i) \\ & + \frac{1}{N_E} \Sigma_{e \in E_i} s(e, cs_i) \end{aligned} \tag{4}$$

Finally, the final reward of the generated CS cs_i can be defined as follows:

$$r(cs_i) = \gamma r_{fact}(cs_i) + (1 - \gamma) r_{faith}(cs_i) \tag{5}$$

where $0 < \gamma < 1$ is the balancing factor.

3.5.3 Reinforcement Tuning

This part aims to optimize the CS generator through reinforcement learning to improve the quality of generated responses. Reinforcement learning has proven to be an effective approach to fine-tuning LLMs to extract complex, useful behaviours from their pre-trained weights (Xu et al., 2024). Existing research (Tian et al., 2023; Yue et al., 2024) indicates that reinforcement learning with proximal policy optimization (PPO) (Schulman et al., 2017) or direct preference optimization (DPO) (Rafailov et al., 2023) can encourage greater factuality and faithfulness in LLMs. In this section, we apply the PPO-based reinforcement learning with a reward $r(\cdot)$ to fine-tune the CS generator. The actor CS generator π_θ is trained to maximise the formulation:

$$\mathbb{E}_{(x,c,E) \sim \{x_i, c_i, E_i\}_{i=1}^{N}, y \sim \pi_\theta(y|\cdot)} [r(x, c, E, y) - \beta D] \tag{6}$$

$$D = \mathbb{D}_{\text{KL}} \left(\pi_\theta(y \mid x, c, E) \| \pi_\phi(y \mid x, c, E) \right) \tag{7}$$

where x is the HS, c is the claim, E is the evidence, π_ϕ is the reference policy. D is the KL divergence term to prevent optimization instability or overoptimization (Gao et al., 2023). β is a hyperparameter to regularize the output difference between π_θ and π_ϕ. During training, we initialize the π_θ and π_ϕ using the weights from the π_{SFT}. Finally, we use the reinforcement-tuned actor model π_θ as the final model for generating CS.

4 Experiments

4.1 Datasets

We conducted experiments on three CS generation datasets, focusing solely on the HS instances from these datasets to generate CS. Detailed information is as follows:

- **CONAN** (Chung et al., 2019) is a large-scale, multilingual resource designed to combat online hate speech through expert-generated counter-narratives. It includes 4,078 pairs of hate speech and counter-narratives in English, French, and Italian, collected by over 100 trained operators from NGOs.
- **MTCONAN** (Fanton et al., 2021) includes 5,000 hate speech and counter-narrative pairs in English, generated using a human-in-the-loop methodology. It covers multiple hate targets such as the disabled, Jews and LGBT+.
- **MTKGCONAN** (Chung et al., 2021) includes pairs of hate speech and counter-narratives from the CONAN dataset, enhanced with external

LLM	Method	CONAN					MTCONAN					MTKGCONAN				
		FA	FH	FC	FE	D	FA	FH	FC	FE	D	FA	FH	FC	FE	D
GPT-3.5	IOP	0.339	0.712	——	——	0.542	0.341	0.701	——	——	0.510	0.371	**0.793**	——	——	0.656
	CoT	0.341	**0.754**	**0.730**	**0.630**	0.714	0.371	**0.749**	**0.712**	0.561	**0.778**	0.365	0.754	**0.721**	**0.589**	0.745
	CoTR	**0.382**	0.704	0.707	0.620	**0.719**	**0.394**	0.720	0.701	**0.564**	0.756	**0.385**	0.763	0.706	0.578	**0.769**
GLM4	IOP	0.331	0.690	——	——	0.604	0.332	0.731	——	——	0.542	0.351	0.767	——	——	0.724
	CoT	0.312	0.721	0.670	0.641	0.768	0.354	0.763	0.631	0.575	**0.753**	0.331	0.764	0.697	0.590	0.759
	CoTR	0.376	0.777	0.654	0.621	**0.742**	0.381	0.799	0.642	0.569	0.740	0.383	**0.819**	0.657	0.601	**0.796**
	F^{2}RL	**0.402**	**0.815**	**0.729**	**0.649**	0.738	**0.421**	**0.817**	**0.711**	**0.643**	0.753	**0.431**	0.812	**0.731**	**0.629**	0.785
Qwen1.5	IOP	0.334	0.754	——	——	0.745	0.301	0.741	——	——	0.702	0.345	0.791	——	——	0.796
	CoT	0.321	0.767	0.609	0.621	**0.802**	0.323	0.803	0.594	0.574	0.760	0.320	0.821	0.654	0.591	0.793
	CoTR	0.352	0.793	0.617	0.604	0.779	0.363	0.799	0.601	0.553	**0.776**	0.388	0.819	0.617	0.588	**0.811**
	F^{2}RL	**0.413**	**0.814**	**0.712**	**0.618**	0.770	**0.431**	**0.821**	**0.714**	**0.637**	0.763	**0.401**	**0.821**	**0.706**	**0.631**	0.784
Llama3	IOP	0.309	0.761	——	——	0.738	0.371	0.762	——	——	0.717	0.302	0.802	——	——	0.780
	CoT	0.319	0.781	0.642	0.593	**0.790**	0.361	0.794	0.661	0.579	0.777	0.329	0.822	0.651	0.621	0.808
	CoTR	0.369	0.752	0.663	0.610	0.763	0.377	0.756	0.655	0.567	0.765	0.339	0.812	0.668	0.589	**0.815**
	F^{2}RL	**0.401**	**0.829**	**0.713**	**0.628**	0.755	**0.416**	**0.822**	**0.717**	**0.621**	0.770	**0.417**	**0.823**	**0.721**	**0.632**	0.793

Table 1: Counterspeech Generation Results. The highest scores of each model are in bold. The overall highest scores are highlighted in red. IOP method generates CS directly from HS, thus we do not evaluate its FC and FE.

knowledge from WikiText-103 and Newsroom. It includes 196 pairs of HS and counter-narratives targeting various hate categories.

4.2 Competing Methods

We selected four LLMs as generators: GPT-3.5, GLM4 (Zeng et al., 2022), Qwen1.5 (Bai et al., 2023), and Llama3 (AI@Meta, 2024). To better understand the effectiveness of each design, we compare three different methods for generating claim-guided and evidence-supported CS, namely:

- **Input-Output Prompt (IOP)**: We designed a simple prompt template that instructs the LLM to directly generate CS with factual evidence support.
- **Chain of Thought (CoT)**: We first use an LLM to generate the claim. Next, we generate evidence based on the claim and the model's parametric knowledge. Finally, we generate the final CS using the claim and the evidence.
- **Chain of Thought with Retrieval (CoTR)**: Different from CoT, we use an external knowledge base to retrieve evidence.

As for F^{2}RL, the method introduced in this paper, we first generate the claim and retrieve supporting evidence. Then, we use an LLM, which has been fine-tuned with supervised fine-tuning and factuality and faithfulness reinforcement learning, as the generator.

4.3 Implementation Details

For the version of LLMs in our experiments, we utilized *gpt-3.5-turbo-0125*, *glm-4-9b-chat*, *Qwen1.5-7B-Chat*, and *Llama-3-8B-Lexi-Uncensored*[*]. These models are used simultaneously for generating claims, queries, and CS. For the claim generation, We generate 5 claims each time. We use the same LLM as the claim generator to act as the voting agent. For evidence retrieval, we generate 5 queries and select the top 5 relevant documents during the retrieval process. We use the off-the-shelf Contriever-MS MARCO as the retriever and BGE M3 as the reranker. We train the BGE M3 model for 5 epochs with a learning rate of 1e-5. During the supervised fine-tuning phase, We split the dataset into 50% for training, 25% for validation, and 25% for testing. Since the dataset lacks ground truth for claim-guided evidence-supported CS, we used the state-of-the-art GPT-4o model to generate it for each HS in the training set. Five volunteers are hired to manually verify and revise the generated ground truth. We perform instruction fine-tuning on the training set using LoRA. During the reinforcement learning phase, we used the TRL[†] library. We adopted PPO with a learning rate of 3e-5 and set the initial KL regularization to 0.2. The training consists of 3 epochs, with a batch size of 16, and parameters are updated after 4 gradient accumulation steps. We conducted all the experiments with Nvidia A800 GPUs.

4.4 Evaluation Metrics

To evaluate the factuality and faithfulness of the generated CS, We follow the evaluation from (He

[*]We also experimented with the Meta-Llama-3-8B-Instruct model. However, due to its strict alignment protocols, it frequently refuses to generate CS in response to hate speech.

[†]https://github.com/huggingface/trl

et al., 2023; Yue et al., 2024) we use four metrics which are factuality (**FA**), faithfulness to Hate Speech (**FH**), faithfulness to Claim (**FC**) and faithfulness for Evidence (**FE**), and Diversity (**D**).

FA can be calculated by equation 3, FH, FC, and FE can be calculated from the individual terms in Equation 4. The diversity (Wang and Wan, 2018) of CS $\hat{y}_i$ in a collection of generated CS $\hat{Y}_i$ is defined using the following formula:

$$\text{Diversity}\left(\hat{y}_i\right) = 1 - \max\left\{\varphi\left(\hat{y}_i, \hat{y}_j\right)\right\}_{j=1}^{j=|\hat{Y}|, j\neq i} \tag{8}$$

4.5 Main Experimental Results

We report the main experimental results of F^{2}RL on three benchmark datasets in Table 1. We draw the following observations. **(1)The F^{2}RL method can improve the factual correctness of CS.** From the experimental results in Table 1, it can be seen that F^{2}RL achieved a 3% to 6% improvement in the FA metric across all three datasets. **(2) The F^{2}RL method generates CS that more strongly opposes hate speech.** For instance, the highest FH across all three datasets is achieved by Llama3-F^{2}RL. Additionally, for various LLMs, F^{2}RL achieved higher FH scores in most cases compared to other baselines. **(3) The F^{2}RL method demonstrates greater consistency with claims and makes better use of existing evidence.** For example, in the MTCONAN dataset, Llama3-F^{2}RL improved the FC metric by approximately 6% compared to Llama3-CoTR. Additionally, GLM4-F^{2}RL enhanced the FE metric by about 7% compared to GLM4-CoT. **(4) F^{2}RL method demonstrates generalizability and effectiveness across different LLMs.** In experiments, the F^{2}RL consistently outperforms other baselines on various LLMs. **(5) The F^{2}RL method leads to a decrease in the diversity of generated CS.** Compared to Llama3-CoTR, Llama3-F^{2}RL shows a slight decrease in diversity metrics across all three datasets. We attribute this decrease in diversity to three potential factors: (i) specific sentence structures: by observing the generated CS, it is evident that LLMs trained with F^{2}RL tend to use certain recurring sentence structures. For example, these models frequently begin with a claim, followed by a transition using phrases such as "in fact," "for example," or "for instance," before listing supporting facts; (ii) single source retrieval: the retrieval sources used in this study are limited to Wikidata. While this ensures the correctness of retrieved information, it restricts the diversity of the generated counterspeech; (iii) reinforcement learning training: reinforcement learning training encourages the model to rely more heavily on retrieved information, which in turn reduces the diversity and creativity of the generated text. However, For evidence-driven counterspeech tasks, factual accuracy outweighs diversity (Chung et al., 2021). The goal of counterspeech is to correct harmful statements, so providing accurate and reliable information is crucial for credibility. Although creativity may boost engagement, it also increases the risk of inaccuracies, potentially reducing effectiveness.

4.6 Ablation Study

Methods	FA	FH	FC	FE
F^{2}RL	0.40	0.83	0.71	0.63
w/o r_{fact}	0.37 ↓0.03	0.84	0.70	0.62
w/o r_{faith}	0.41	0.76 ↓0.07	0.67 ↓0.04	0.60 ↓0.03
w/o FH	0.40	0.75 ↓0.08	0.71	0.64
w/o FC	0.41	0.83	0.66 ↓0.05	0.62
w/o FE	0.40	0.82	0.72	0.59 ↓0.04

Table 2: Experimental results of ablation study.

In this section, we aim to explore how much each reward part contributes to the performance during the reinforcement learning phase. We consider the following variants of F^{2}RL: (1) **F^{2}RL w/o** r_{fact}: During the reinforcement learning process, we remove the reward calculation for factuality. (2) **F^{2}RL w/o** r_{faith}: we remove the reward calculation for faithfulness. (3) **F^{2}RL w/o FH**: We remove the reward calculation for faithfulness to hate speech. (4) **F^{2}RL w/o FC**: We remove the reward evaluation for faithfulness to claim. (5) **F^{2}RL w/o FE**: We remove the reward evaluation for faithfulness to evidence.

We evaluate the aforementioned variants using the Llama3 model on the CONAN dataset. The experimental results are shown in the Table 2. We find that the removal of r_{fact} causes a 3% decrease in FA, which verifies the effectiveness and significance of the factuality reward feedback. When r_{faith} is removed, the faithfulness of the generated CS in all aspects decreases. We also conducted a more fine-grained study, and when different parts of r_{faith} were removed, their corresponding metrics showed varying degrees of decrease. This indicates that each component of the faithfulness reward feedback plays a crucial role in maintaining the faithfulness of the generated CS.

4.7 Human evaluation

		FA	FH	FC	FE
GPT-3.5	CoT	2.97	3.71	3.46	2.78
	CoTR	3.56	3.69	**3.74**	2.67
Llama3	CoT	2.67	3.68	3.24	2.56
	CoTR	3.34	3.54	3.35	2.76
	F^{2}RL	**3.61**	**3.74**	3.68	**2.89**

Table 3: Human evaluation results.

In human evaluation, we randomly sample 100 CS generated by GPT-3.5 and Llama3 from the CONAN dataset. Given the HS, claim, evidence, and the generated CS, we recruit five annotators (majority rule) to assign a score from 1 to 5 (1: not at all, 3: OK, 5: very good) to the generated CS based on the aspects of factuality, faithfulness for HS, claim and evidence. The age distribution of the 5 volunteers is between 20 and 30 years old. Among them, there are 4 males and 1 female. Two are PhD students, and three are master students. The four aspects are (1) Factuality: annotators can use various retrieval tools (e.g., Bing, Google) to verify the factual correctness of the CS. (2) faithfulness to HS: Whether the CS explicitly refutes the HS. (3)faithfulness for the claim: whether the CS is consistent with the claim. (4) faithfulness for evidence: whether the CS sufficiently utilizes the evidence. The human ratings results are listed in Table 3. We can observe that F^{2}RL is better than other baselines and achieves performance comparable to GPT-3.5, which is similar to the results of the automatic evaluation in Table 1.

4.8 Impact of the number of Evidence

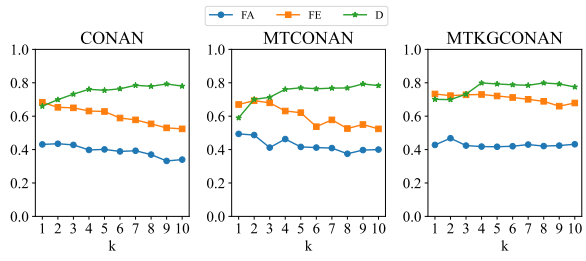

Figure 3: Experimental results of different values of k.

To analyze the impact of using different values of the number of evidence k on performance, we conducted experiments using Llama3-F^{2}RL on three datasets, with results shown in Figure 3. Firstly, as the amount of evidence increases, FA tends to decrease. This indicates that too much evidence in the prompt is more likely to cause hallucinations in LLMs, leading to factually incorrect CS. More evidence also increases noise, reducing the evidence utilization rate and lowering the average FE. When more evidence is provided, LLMs tend to choose and cite only a few pieces rather than all of them. This results in increased diversity in the generated CS, as LLMs have more options to support the claim.

5 Conclusion

In this paper, we introduce F^{2}RL, a factuality and faithfulness reinforcement learning framework for claim-guided, evidence-supported CS generation. To ensure that CS has a distinct argument and a clear evidence structure, we propose a claim-guided pipeline that first generates a counter-claim, then retrieves relevant evidence, and finally generates the CS. To enhance the coherence and evidence connection of the CS, we design a coarse-to-fine evidence retrieval method. This method first generates broad queries to ensure the diversity of evidence, then carefully reranks the results to ensure their relevance to the claim. To improve the factuality and faithfulness of the CS, we propose a PPO-based reinforcement learning approach with a triplet-based factuality reward model and a multi-aspect faithfulness reward model. Experimental results show that our method outperforms competitive baselines in terms of factuality and faithfulness. In the future, we aim to explore LLM-based multi-agent learning methods to further improve the generation of CS.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments and helpful suggestions. This work was supported by the National Natural Science Foundation of China (No.62172428).

Limitations

There are three limitations of this work: (1) Firstly, it is important to note that the generated CS by our model cannot completely eliminate offensive and toxic language. However, we have made improvements to the prompt template and generation process to generate CS with lower toxicity. (2) Secondly, it is worth mentioning that the generated CS cannot be directly posted on real-world

social media platforms. Instead, a further step is required where volunteers carefully review and verify the generated CS before posting. This automated generation process has proven to be a significant time-saving factor compared to the manual creation of CS. (3) Thirdly, the counterspeech generated by our method may still contain factual errors. While our approach aims to minimize the likelihood of such errors, it cannot guarantee their complete elimination.

Ethics Statement

The goal of this work is to diminish hate intensity and promote a harmonious communication environment. However, we acknowledge the sensitive nature of combating online hate speech. Our research may give rise to some ethical and moral considerations. Nevertheless, we are confident that our work aligns with the ethical policy established by EMNLP.

References

AI@Meta. 2024. Llama 3 model card.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Considerations for successful counterspeech. *Dangerous speech project*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding counterspeech for online harm mitigation. *CoRR*, abs/2307.04761.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2819–2829. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 899–914. Association for Computational Linguistics.

Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. Online hate speech victimization: consequences for victims' feelings of insecurity. *Crime Science*, 13(1):4.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,*

ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 3226–3240. Association for Computational Linguistics.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5792–5809. Association for Computational Linguistics.

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, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2698–2709. ACM.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *CoRR*, abs/2404.10981.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tang, Haizhou Wang, and Wenxian Wang. 2023. RAUCG: retrieval-augmented unsupervised counter narrative generation for hate speech. *CoRR*, abs/2310.05650.

Hao Lang, Fei Huang, and Yongbin Li. 2024. Fine-tuning language models with reward learning on policy. *CoRR*, abs/2403.19279.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *CoRR*, abs/2203.05227.

John T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6516–6528. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5157–5163. ijcai.org.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18990–18998. AAAI Press.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *CoRR*, abs/2311.08401.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,

Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Jeremy Waldron. 2012. *The harm in hate speech*. Harvard University Press.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 9131–9142. ELRA and ICCL.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023. Boosting language models reasoning with chain-of-knowledge prompting. *CoRR*, abs/2306.06427.

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4446–4452. ijcai.org.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? A comprehensive study. *CoRR*, abs/2404.10719.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: rank responses to align language models with human feedback without tears. *CoRR*, abs/2304.05302.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. pages 5628–5643.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *CoRR*, abs/2402.19473.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 134–149. Association for Computational Linguistics.

A Analysis of Typical Argumentative Sentence Structures

We analyzed the typical argumentative sentence structures in claim-guided, evidence-supported CS. We found that they generally consist of three stages: first, refuting the erroneous viewpoints of the hate speech; second, presenting benign statements about the hate target; and finally, providing evidence. We listed some examples in the Table 4.

B Case Study

We provided counterspeech examples generated by GPT-3.5 and Llama3 to compare the quality of generation. See Table 5 for the examples. Through comparison, we can see that the counterspeech generated by GPT-3.5 (IOP) lacks specific evidence and is more general. Both GPT-3.5 (CoT) and GPT-3.5 (CoTR) provide factual evidence as support, with GPT-3.5 (CoTR) offering more detailed and in-depth arguments. Compared to Llama3 (CoTR), Llama3 (F^2RL) discusses the claim in more detail and provides multiple logically connected pieces of evidence.

C Prompt Template

C.1 Claim Generation Prompt

Generate {claim num} claims to refute the statement:{hate speech} These claims should be well-structured and provide a strong foundation for retrieving factual knowledge to support them. You only need to generate short, direct claims without providing facts. List each counter-claim as follows:

[claims]:

1. [First counter-claim]
2. [Second counter-claim]
3. [Third counter-claim]

Counter hate speech	State a claim	Provide evidence
[**HT**] are not [**TS**]	In fact, they are [**BS**].	For example, [**E**].
Contrary to the statements that [**HT**] are [**TS**]	Studies have shown that they are [**BS**]	For instance, [**E**].
It's incorrect to say that [**HE**] are [**TS**]	Data indicates that they are actually [**BS**].	Consider [**E**].
The notion that [**HT**] are [**TS**] is false.	Evidence suggests that they are [**BS**].	To demonstrate, [**E**].
The stereotype that [**HT**] are [**TS**] is unfounded.	In reality, they are [**BS**].	Such as, [**E**].

Table 4: Typical argumentative sentence structures of claim-guided evidence-supported CS. **HT** stands for hate target, **TS** stands for toxic statement, **BS** stands for benign statement, and **E** is evidence.

Hate Speech	*Not all Muslims are terrorists, but all terrorists are Muslims.*
GPT-3.5 (IOP)	This statement is inaccurate. In reality, the majority of global terrorist attacks are not committed by Muslims. Terrorism involves individuals and groups from various backgrounds, religions, and ideologies worldwide. The definition of terrorism is based on actions, not religious beliefs. Let's not judge entire groups based on narrow viewpoints.
GPT-3.5 (CoT)	This statement is misleading. Global terrorism data clearly shows that most terrorist attacks are not committed by Muslims. Historical incidents, like the Oklahoma City bombing by Timothy McVeigh, highlight that terrorism isn't exclusive to Muslims. Terrorism is defined by actions, not religious identity. Let's avoid generalizations and focus on facts.
GPT-3.5 (CoTR)	Terrorism is not exclusive to any particular religion or belief system. While it is true that some terrorist acts have been carried out by individuals claiming to be Muslims, it is important to recognize that terrorism is a tactic used by individuals of various backgrounds and ideologies. For example, Jamaat ul-Fuqraa is an Islamic terrorist organization with members who have planned acts of violence, but they are not the only group engaging in terrorism. Religious terrorism, including acts committed by individuals of different faiths, has been on the rise since 1980. It is crucial to understand that terrorism is a complex issue influenced by various factors, and attributing it solely to one religion is misleading and harmful.
Llama3 (CoTR)	The claim that "all terrorists are Muslims" is false. Evidence includes groups like Jamaat ul-Fuqra, an Islamic terrorist organization mostly African-American, responsible for violence like assassinations. Religious terrorism isn't limited to Islam; e.g., the 1995 Oklahoma City bombing stemmed from anti-government and anti-abortion beliefs. Experts like Robert Pape and Terry Nardin note terrorism's broader motivations beyond religion.
Llama3 (F^2RL)	Terrorism isn't tied to any specific religion or belief system. While some attacks involve individuals identifying as Muslim, it's not representative of all terrorists. Evidence includes groups like Jamaat ul-Fuqra, an Islamic terrorist organization mostly composed of African-Americans, responsible for acts of violence like assassinations. Religious terrorism isn't limited to Islam; for example, the 1995 Oklahoma City bombing was carried out by Timothy McVeigh, motivated by anti-government and anti-abortion beliefs. Furthermore, the 2011 Norway attacks by Anders Behring Breivik were driven by far-right extremist views, not Islam. Experts like Robert Pape and Terry Nardin note that terrorism's broader motivations often include political, ideological, and social factors beyond religion. This demonstrates that terrorism is a complex, multi-faceted issue that transcends any single religious or belief system.

Table 5: Evidence-supported Counterspeech generated by CPT-3.5 and Llama3.

C.2 Voting prompt

Given an hate speech: "{hatespeech}" and several counter-claims, decide one that can most effectively counter the hate speech. Conclude in the last line "The best claim is [s]", where s the integer id of the choice.

claims: {claims}

C.3 Query Generation Prompt

Given a hateful statement and a corresponding counter-claim, your task is to generate {query num} search queries to retrieve evidence from Wikipedia that supports the counter-claim. The query should be precise and relevant to ensure the retrieval of strong factual evidence. Here is the hateful statement: "{hatespeech}" Here is the counter-claim: "{claim}" Please generate a search query for Wikipedia to find evidence supporting the counter-claim. List each query as follows: [queries]: 1. [the First query] 2. [the Second query] 3. [the Third query] 4. [the Second query] 5. [the Third query]

C.4 Counterspeech Generation Prompt

You are a seasoned volunteer dedicated to countering hate speech on social media. Given a claim and relevant evidence of each claim, your task is to generate a counterspeech. The Counterspeech needs to first state the claim and then provide evidence to support the claim. The counterspeech should be effectively refute the hatespeech. You must give me the counterspeech as following format: [Counterspeech]: ""

Here is the hate speech: "{hatespeech}" Here is the counter-claim: "{claim}" Here is the evidence: "{evidence$_s tr$}"