Integrating Argumentation and Hate-Speech-based Techniques for Countering Misinformation

Sougata Saha[1,2] **and Rohini Srihari**[1]
[1]State University of New York at Buffalo
Department of Computer Science and Engineering
[2]MBZUAI,
[1]{sougatas, rohini}@buffalo.edu, [2]sougata.saha@mbzuai.ac.ae

Abstract

The proliferation of online misinformation presents a significant challenge, requiring scalable strategies for effective mitigation. While detection methods exist, current reactive approaches, like content flagging and banning, are short-term and insufficient. Additionally, advancements like large language models (LLMs) exacerbate the issue by enabling large-scale creation and dissemination of misinformation. Thus, sustainable, scalable solutions that encourage behavior change and broaden perspectives by persuading misinformants against their viewpoints or broadening their perspectives are needed. To this end, we propose persuasive LLM-based dialogue systems to tackle misinformation. However, challenges arise due to the lack of suitable datasets and formal frameworks for generating persuasive responses. Inspired by existing methods for countering online hate speech, we explore adapting counter-hate response strategies for misinformation. Since misinformation and hate speech often coexist despite differing intentions, we develop classifiers to identify and annotate response strategies from hate-speech counter-responses for use in misinformation scenarios. Human evaluations show a 91% agreement on the applicability of these strategies to misinformation. Next, as a scalable counter-misinformation solution, we create an LLM-based argument graph framework that generates persuasive responses, using the strategies as control codes to adjust the style and content. Human evaluations and case studies demonstrate that our framework generates expert-like responses and is 14% more engaging, 21% more natural, and 18% more factual than the best available alternatives.

1 Introduction

Misinformation is the unintentional falsification of information and can be a life-threatening menace (Galvão, 2021) on online social media platforms. The public availability of advanced technologies such as LLMs, enables the rapid creation of false information at scale, and the ease of access to the web across all demographics aids its dissemination (Wilson and Maceviciute, 2022; Pan et al., 2021; Allcott et al., 2019), causing an uptake in the acceptance of fake news. While most means of combating misinformation (Collins et al., 2021) resort to myopic approaches like content flagging and banning, they do little to change the perception of the misinformant. Although such reactive measures reduce the spread of false information, they are momentary, as the misinformant still holds the incorrect perception and is likely to re-share the false information again. Hence, we need scaleable solutions that attempt perception change, leading to a lasting reduction in the spread of false news.

As per Micallef et al. (2020a), 96% of counter-misinformation responses are by other (non-expert) social media users, which effectively curbs misinformation (Walter et al., 2021, 2020; Walter and Murphy, 2018) and reduces misperceptions (Bode and Vraga, 2021; Colliander, 2019; Friggeri et al., 2014; Seo et al., 2021; Wijenayake et al., 2020) across topics (Bode and Vraga, 2015; Bode et al., 2020; Vraga and Bode, 2018, 2021; Bode and Vraga, 2018; Vraga and Bode, 2020), platforms, and demographics (Vraga et al., 2022a,b, 2020). Although scalable, unlike experts, most non-expert user responses are rude and use unverified evidence (Micallef et al., 2020b; He et al., 2023), propelling mistrust (Flekova et al., 2016; Thorson et al., 2010) and further agitation (Cheng et al., 2017; Kumar et al., 2018; Masullo and Kim, 2021). On the other hand, expert-curated responses are factually consistent, more informative, and more effective. However, the expert responses are mostly template-like and generic, making them less engaging and persuasive. Furthermore, there are far fewer experts than non-experts, which makes it hard to keep up with the volume of misinformation. Hence, we propose persuasive dialogue systems as a comple-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11109–11124

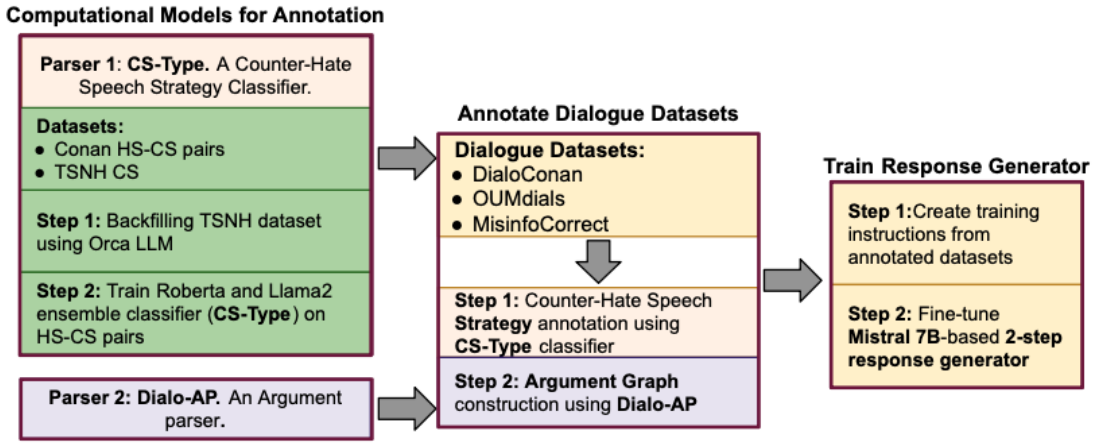

Figure 1: High-level architecture of the proposed solution.

mentary solution that combines the naturalness, variety, and scalability of non-expert-based counter-responses with the informativeness and politeness of expert-based responses. However, implementing such systems requires appropriate dialogues with well-defined counter-misinformation response strategy annotations as training data, which are lacking. Furthermore, although persuasion is theoretically well studied, computational frameworks are limited, which we address through our solution.

Online misinformation and hate speech are often interwoven in practice (Kim and Kesari, 2021; Cinelli et al., 2021; Kim and Kesari, 2021), where either one leads to another or is co-existent in posts. However, unlike misinformation, formal definitions of counter-hate response strategies exists (Benesch et al., 2016; Chung et al., 2019; Mathew et al., 2019). Therefore, we adapt those strategies to respond to misinformation posts. We train computational models to identify counter-hate response strategies and analyze the utility of adapting such strategies to combating misinformation. We computationally annotate such strategies in diverse datasets comprising hate speech dialogues (Bonaldi et al., 2022), polarized factual discussions (Farag et al., 2022), and misinformation (He et al., 2023), yielding a silver annotated misinformation and hate-speech dialogue corpus[1]. Furthermore, we represent dialogue as an argument graph (Peldszus and Stede, 2013) comprising claims and premises (Van Eemeren et al., 2002; Besnard and Hunter, 2008; Van Eemeren, 2015) from the dialogue turns, and formalize persuasion as a series of decision-making in the graph. As a response, the generator first determines the contextual argument components to attack/support and the hate speech-based response strategy. Next, it generates the response claims and premises, followed by the final response text. Such a formalism fosters controllability and explainability of the generated responses. Figure 1 illustrates our solution at a high level. Our contributions are:

- We adapt counter-hate speech strategies for responding to misinformation.
- We share the silver-annotated MisinfoCorrect (He et al., 2023) dataset comprising COVID-19 vaccine misinformation and response pairs with the counter-response strategies.
- We represent dialogue as an argument graph and share a framework for persuasion.
- We implement the framework using controllable LLMs that strategize and generate responses to hate speech and misinformation.

2 Related Work

Computational Persuasion: Hunter (2018) defines computational persuasion as the study of formal models of dialogues involving arguments, counterarguments, and strategies, and outlines a framework for fostering behaviour change. Meade (2021) discusses the different persuasive strategies evident from daily life, whereas Cialdini and Cialdini (2007) studies persuasive strategies in sales and marketing. Hornsey and Fielding (2017) discusses attitude roots as the underlying motivations in people and developed a "jiu jitsu" model of persuasion. Ruiz Dolz (2023) devised an argument-based computational persuasive framework. Al-

[1]Dataset link:https://github.com/sougata-ub/MisInfoCorrected.git

hindi (2023) used argument structure and quality for improving fact checkers.

Controllable Argument Generation: Controllable argumentation has benefited from the advancements in language modeling. Keskar et al. (2019) regulated language models using control codes. Hua et al. (2019); Hua and Wang (2018) experimented with factual counter-argument generators that can be controlled for the style. Schiller et al. (2021) introduced Arg-CTRL: a language model for generating sentence-level arguments using topic, stance, and aspect-based control codes. Khatib et al. (2021) constructed argumentation-related knowledge graphs and used them to control argument generation. Syed et al. (2021) used control codes in extractive and abstractive models for generating conclusions from arguments. Chakrabarty et al. (2021) controllably re-framed polarized arguments to reduce their fear quotient. Saha and Srihari (2023) used Walton's argument scheme-based control codes for generating factual arguments around polarized topics. As a solution, Alshomary et al. (2021) ranked argument premises by their strength and attacked the weakest premise. Saha and Srihari (2024) used features based on Walton's argument schemes, speech acts, personality traits, and human values for controlling persuasive counter-responses.

Counter Hate-Speech and Misinformation: Neural approaches like CounterGeDi (Saha et al., 2022a), Toxicbot (de los Riscos and D'Haro, 2021), and others (Kiritchenko et al., 2021; Windisch et al., 2021; Zhu and Bhat, 2021), has advanced counter-hate speech generation. Advancements have also been made in knowledge-based and multi-lingual models such as Chung et al. (2021a) and Chung et al. (2020). However, multi-turn hate speech-based dialogues are lacking. Conan (Chung et al., 2019), offers pairs of hate speech and counter-narratives in English, French, and Italian, primarily addressing Islamophobia. Building upon Conan, Multi-Target Conan (Fanton et al., 2021) broadened the scope to include English examples covering various hate targets. Moreover, Chung et al. (2021b) introduced a subset of Multi-Conan examples grounded in knowledge. Additionally, Furman et al. (2022) curated responses to hateful comments from the Hateval corpus (Basile et al., 2019) and enhanced it with annotations derived from Wagemann's periodic table of arguments (Wagemans, 2016). Furthermore, Chasm (Ashida and Komachi, 2022) utilized LLMs to produce a synthetic, quality-controlled dataset for counter-hate speech, offering additional annotations conducive to counter-narrative evaluation research. Lastly, Bonaldi et al. (2022) introduced DialoConan, a dataset consisting of multi-turn dialogues synthetically generated and quality-controlled, depicting interactions between a hater and an NGO operator.

Unlike hate speech, frameworks and datasets for counter-responses to misinformation are still lacking. To address this, He et al. (2023) created MisinfoCorrect dataset comprising COVID-19 misinformation Tweets and counter-responses by experts and non-experts, and trained a factual response generator to refute false information politely. However, the dataset has a few limitations. First, it only contains misinformation posts and response pairs, making it non-conducive to lengthy conversations. Second, the expert-curated responses are usually informative and generic and do not contain diverse response styles, unlike the non-experts. Hence, we aim to train generators capable of informative and engaging responses conducive to dialogues.

3 Adapting Hate Speech Response Strategies to Misinformation

Counter-responses strategies to online hate speech are well studied. Benesch et al. (2016) analyzed approximately 1M controversial Tweets about Islamophobia, LGBTQ, Race, Refugees, Women, and Politics and defined a taxonomy of nine strategies of counter-responses to hate comments, detailed in Figure 8 (Appendix A). Using Benesch et al. (2016)'s strategy definitions, Mathew et al. (2019) shared an annotated dataset of 7K examples (TSNH) containing counter-responses to YouTube comments targetting Jews, LGBT+, and People of Color (POC). Chung et al. (2019) released Conan, a curated dataset of 6.5K hate-speech and counter-response pairs about Islamophobia, and annotated the responses with Benesch et al. (2016)'s response strategy definitions. Consolidating the Conan and TSNH examples, we train an ensemble of LLM-based classifiers for identifying counter-response strategies from given hate speech and counter-response pairs. However, since the TSNH dataset does not include the original hate comment of a counter-response, we first train an LLM-based generator for synthetically generating (backfilling) hate comments from counter-responses and then train the LLM-based counter-response type (**CS-Type**) ensembled classifier on the Conan and syn-

thetically enriched TSNH dataset, and use it to annotate the counter-responses to misinformation posts in the MisinfoCorrect dataset.

3.1 Backfilling Hate Comments

Identifying the response strategy without the original comment is difficult. For example, the response "Well, I do find similarities" points out hypocrisy if it responds to the comment "Unlike us, X is a hateful and non-tolerant religion" and denounces it if it responds to "People who are X are different from Y". Since the TSNH dataset does not include the original hate comments, we use orca_mini_v3_7b, the Llama2-7b-based version (Touvron et al., 2023) of Orca LLM (Mukherjee et al., 2023; Mitra et al., 2023) containing 7 billion parameters for synthetically generating proxy hate comments of counter-hate responses. We train Orca on 35,000 examples from the Conan, MultiConan, DialoConan, and ASFoCoNG datasets using three types of instructions: (i) 15,000 instructions from all four datasets for generating the counter-hate response to hate speech. (ii) 15,000 instructions for generating the preceding hate speech of a counter-hate response (**backfilling**). (iii) 5,000 instructions from DialoConan for generating the **follow-up** response by the hateful speaker in response to the counter-hate response. Figure 6 (Appendix A.2) illustrates the instructions, and Table 5 (Appendix A) describes each dataset in detail. Although we only intend to backfill the hate comments of the TSNH dataset, following the literature (Taori et al., 2023; Zhang et al., 2023), we use diverse instructions to enhance model robustness. We randomly create training and testing sets of 33,000 and 2,000 examples and finetune the q and v projection modules of Orca using LoRA (Hu et al., 2021; Mangrulkar et al., 2022). The model is trained for 19 epochs, setting LoRA r and α to 8 and 16, and 3e-4 learning rate.

3.2 Counter-Speech Strategy Classifier

We backfill the TSNH dataset and combine it with the Conan pairs to create a dataset of 13.5K examples comprising hate speech-counter-speech pairs and multi-class labels identifying the counter-speech strategy. We randomly split the dataset as 10K training, 1.5K validation, and 2K test and finetune a Roberta-large (Liu et al., 2019) multi-label classifier to predict the response strategies (or others) given a concatenated hate speech and response input pair. Table 1 (column **Rob-Pairs**) shares the classifier's strategy-wise test F1 score and **frequency**. We also train Roberta on examples comprising Conan pairs and the TSNH original responses without backfilling and share the test F1 score in Table 1 (column **Rob-No Pairs**). We observe that for overall and the top 4 frequent (68%) strategies, finetuning using the backfilled data yields better results.

Id	Strategy	Freq	Rob-Pairs	Rob-No Pairs	Llama-Pairs
1	Presenting facts	23%	**84.7**	83.1	85.9
2	Denouncing	16%	**64.7**	64.5	66.3
3	Hostile language	16%	**77.0**	74.7	76.6
4	Pointing out hypocrisy or contradiction	13%	**64.1**	62.8	69.8
5	Humor and sarcasm	10%	65.0	**67.4**	75.1
6	Positive tone	8%	**64.7**	62.2	62.3
7	Counter question	6%	85.0	**86.5**	88.3
8	Warning of consequences	4%	62.2	**64.8**	65.5
9	Affiliation	4%	**68.4**	68.3	70.6
10	Other	1%	**25.0**	0.0	31.6
11	**Overall**		**Avg** **66.1**	63.4	69.2
			Wtd **72.2**	71.3	74.6

Table 1: Counter-response type prediction F1 scores on the test set. **Avg:** Arithmetic mean, **Wtd:** Arithmetic mean weighted by frequency. For each strategy the best performing Roberta model is highlighted in bold and the best performing overall model is underlined.

Following the superior results using backfilled data, we also finetune Llama-2-7b-chat as a multi-label sequence classifier using LoRA on instructions from the hate speech-response pairs, where an instruction comprises a brief description of the nine strategies, followed by the hate speech-response pair as input. The projection modules of the model's attention heads are trained for ten epochs, setting LoRA r and α values to 8 and 16 and a learning rate of 3e-4. We observe that Llama-2 (Table 1 column **Llama-Pairs**) attains the best overall F1 scores and outperforms Roberta-large across all classes, except for Hostile language and Positive tone. The final classifier (**CS-Type**) ensembles the Roberta and Llama paired variants by thresholding the argmax scores at 0.3 and 0.25.

3.3 Evaluations and Analysis

3.3.1 Human Evaluation

We annotated the counter-responses in the MisinfoCorrect dataset using the CS-Type classifier and evaluated 54 random predictions in Amazon Mechanical Turk (AMT). We provided descriptions of the nine response strategies (and a none option) with comment-response pairs as examples. Each

prediction was evaluated by two evaluators for correctness on a binary scale and paid 1 US Cent per evaluation, adhering to the AMT paying guidelines. The evaluators agreed in 52 out of 54 cases (96%) and disagreed in 2 samples. The model predictions were deemed correct in all cases.

We also employed two Computer Science and Linguistics graduate students and evaluated 50 random predictions following the same AMT guidelines. The evaluators agreed in 43 cases (86%), where 40 (80%) were deemed correct and 3 (6%) incorrect, signifying that the counter-responses can be categorized using the nine strategies. The high inter-annotator agreement and scores further demonstrate that it is possible to correctly adapt the hate-speech-based counter-response strategies to classify counter-misinformation discourse with high accuracy.

3.3.2 Response Strategy Analysis

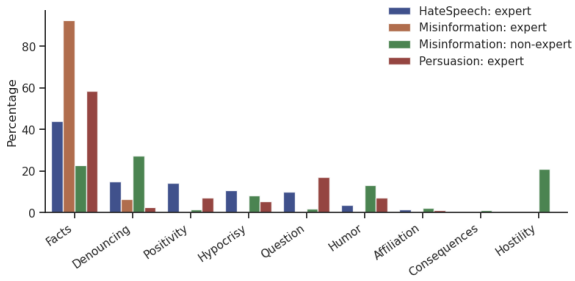

Figure 2: Strategy Comparison.

We also annotate the response strategies in the OUMdials and DialoConan datasets using the CS-Type classifier and compare the strategies implemented by each discourse type. Since the MisinfoCorrect dataset constitutes responses by experts and non-experts, we analyze them independently. Figure 2 plots the distribution of each type of discourse. We observe that (i) All expert-based discourse primarily share facts, where hate speech uses facts in 44% of responses, 59% of cases in persuasion, and 93% in counter-misinformation responses, which is the highest compared to others. Non-expert-based counter-misinformation responses use facts in 23% of cases, making it the lowest among others. (ii) Apart from sharing facts, 15% of counter-hate responses denounce, depicting a reactive intent. Counter-questioning and pointing out hypocrisy is also prevalent in 11% of responses each. (iii) 17% persuasive responses use counter questions and do not denounce, depicting an intent to engage in a dialogue rather than being solely reactive. Persuasion also uses humor and positivity and points out hypocrisy in 7% of cases each, which is lower than hate speech. (iv) Unlike persuasion, 6% expert-based counter-misinformation responses denounce and do not counter-question or use other strategies. Solely using facts makes a response generic and non-engaging and generally does not lead to dialogues, thus limiting their persuasiveness. (v) Confirming He et al. (2023)'s findings, non-expert-based counter-misinformation responses are primarily rude and distinct from the other forms of discourse. They denounce (27%), use hostile language (22%), and are unfit for persuasive response generators. They also use humor (13%) and point out hypocrisy (8%).

Our analysis demonstrates that the distribution of counter-misinformation strategies used in MisinfoCorrect are distinct from other forms of discourse. Furthermore, their effectiveness in reducing misinformation is unmeasured. On the contrary, Benesch et al. (2016)'s and Farag et al. (2022)'s analysis have demonstrated the efficacy of the hate-speech and persuasion-based strategies, which motivates us to use them for responding to misinformation. Furthermore, conditioning language models on diverse strategy-based features enables controllability, where distinct responses can be generated by changing the strategy.

4 Persuasive Response Generator

Here, we experiment with computational models for incorporating the hate speech-based counter-response strategies during response generation such that changing them yields varied responses. Furthermore, we experiment with models that contextually plan the response strategy before generating the response. Since persuasion relies on effective argumentation (Gnamus, 1986; O'Keefe, 2012; Rosenfeld and Kraus, 2016; Roque, 2017), we adopt an argumentation-based framework where we represent a dialogue as an argument graph. The generator sequentially generates a plan determining the contextual graph nodes to attack and support, determines the hate speech-based response strategy to use in the response, generates the appropriate response arguments, and finally combines the arguments as the final response text.

4.1 Dialogue as an Argument Graph

A dialogue consists of multiple alternating interlocutor turns. Each turn, if argumentative, com-

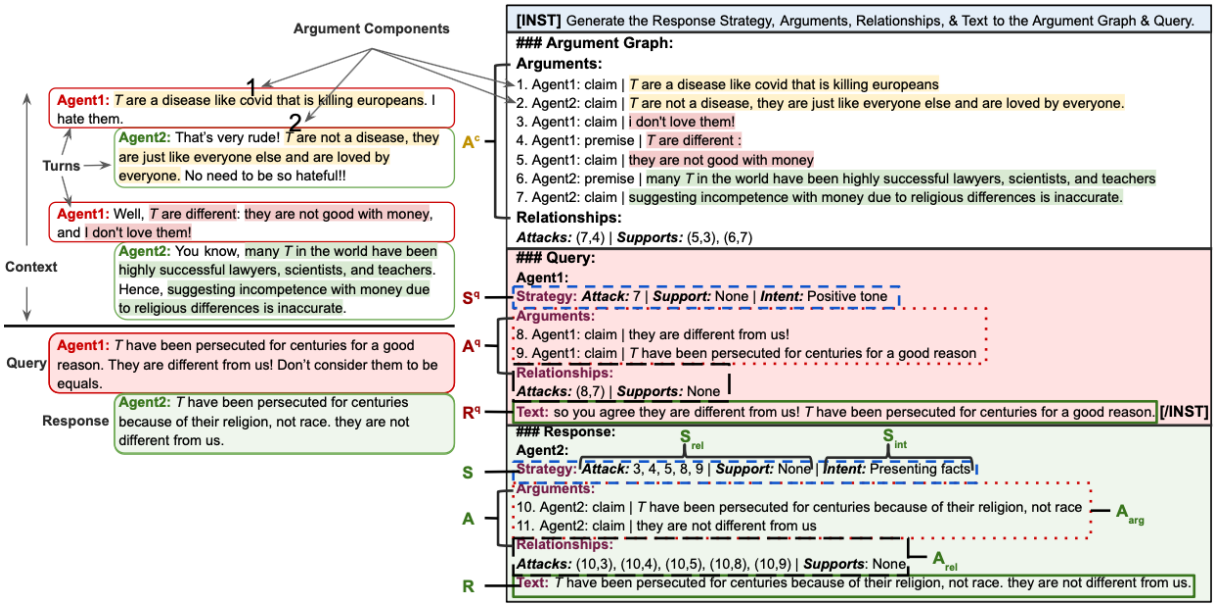

Figure 3: (Left): Sample conversation. (Right): Instruction template from the conversation. Sensitive parts redacted.

prises argument components (**A_arg**) labeled as claims or premises. The turn components support or attack each other (intra) and other contextual components (inter) and form an argument graph ($\mathbf{A}$) when represented as nodes with the support and attack relationships as edges. Such graph-based representations summarize the discourse and find use in discourse analysis and response generation (Chalaguine and Hunter, 2020; Slonim et al., 2021). Figure 3 (left) depicts a multi-turn dialogue between Agents 1 and 2, and (right) illustrates the argument graph representation of the context ($\mathbf{A^c}$). We use Dialo-AP (Saha et al., 2022b) (minimum 0.4/0.1 intra/inter-turn threshold), a dependency parsing-based argument parser, to construct the argument graph of a dialogue and sequentially number each argument component by its order of appearance while re-labeling the major claims as claims.

4.2 Response Generation

We model response generation as a 2-step sequential process consisting of (i) **Response planning/strategy (S):** Plan the (a) Logic ($\mathbf{S_{rel}}$): contextual argument components ($\mathbf{A^c_{arg}}$, $\mathbf{A^q_{arg}}$) to attack and support, and the (b) Strategy ($\mathbf{S_{int}}$): Hate speech-based counter-response type (intent) to exhibit by the response text. (ii) **Realization:** Generate (a) the response argument graph $\mathbf{A}$ consisting of $\mathbf{A_{arg}}$(nodes)- the response arguments and their claim/premise labels, $\mathbf{A_{rel}}$(edges)- the inter and intra-attack and support relationships, and (b) the response text $\mathbf{R}$. Mathematically, let $\mathbf{R^q}$, $\mathbf{A^q}$, $\mathbf{S^q}$ represent the query response, argument graph, strategy, and $\mathbf{A^c}$ represent the context argument graph. Then, the joint probability, $P(R,A,S,Q)=P(R|A,S,Q)*P(A|S,Q)*P(S|Q)*P(Q)$, where $Q=(R^q,A^q,S^q,A^c)$, $A^q=(A^q_{arg}, A^q_{rel})$, $S^q=(S^q_{rel}, S^q_{int})$, $A^c=(A^c_{arg}, A^c_{rel})$, $S^c=(S^c_{rel}, S^c_{int})$

4.3 Training and Experiments

We derive 7,132 examples (6,504 train, 334 dev, and 294 test) from the strategy-annotated DialoConan, OUMdials, and MisinfoCorrect datasets and use the template in Figure 3 to create instructions and limit to 600 tokens. The perpetrator is labeled Agent 1 and the responder is Agent 2.

4.3.1 Training Details

We experiment with Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), an instruction-fine-tuned version of Mistral comprising 7 billion parameters. The model (**dual**) is fine-tuned for two epochs on the consolidated instructions using LoRA. For LoRA, we set r and α to 16 and 32 and train the LM head and the q, k, v, o, gate, up, and down attention head projection modules with a 2.5e-5 learning rate. Greedy decoding is used during inference while limiting the new tokens to 256.

4.3.2 Ablations

We also train the following two ablated model variants without the strategy: (i) **graphText**: Only generate the response arguments, relationships, and text. (ii) **onlyText** (baseline): Only generate the response text. Figure 7 (1-3 Appendix A.2) shares

		Generated vs Gold Text			Generated vs Gold Planning			Generated vs Re-Parsed Planning			
Id	**Variant**	**BLEU**	**ROUGE**	**Rank**	**Attack**	**Support**	**Strategy**	**Attack**	**Support**	**Strategy**	**Type (F1)**
1	dual	3.0	17.0	2.03	59.5	**73.7**	**55.3**	**57.4**	**92.6**	**66.7**	91.4
2	graphText	2.7	16.4	**1.87**	**60.4**	69.5	-	55.8	90.4	-	**94.7**
3	dual+Amp	**3.2**	**17.7**	2.10	60.2	72.7	55.0	51.5	87.5	55.4	88.7
4	onlyText	2.5	16.8								

Table 2: Automatic and human evaluation of the response generator variants. Best scores are highlighted in bold.

the instruction-fine-tuning templates. To further enhance the dual model's argumentation capabilities, we create **dual+Amp**, a fine-tuned version of the dual model on 752 (675 train, 37 eval, 40 test) additional argument mining-based instructions from Ampersand (Chakrabarty et al., 2019) in a multi-task learning framework. Ampersand is an expert annotated dataset that identifies the argumentative components and inter/intra-turn support/attack relationships from 112 Change My View subreddit threads, thus conducive to argument mining tasks from dialogues. Our fine-tuning instructions pertain to generating the response strategy, relationships, and the combined response strategy, arguments, and relationships from an argument graph, as detailed in Figure 7 (4-6 Appendix A.2).

4.4 Evaluations

4.4.1 Evaluating Response Text

Table 2: **Generated vs Gold Text** compares the average BLEU (Papineni et al., 2002) and Rouge-L (Lin, 2004) scores between the original and generated response text using the test set instructions. Since the **dual** and **dual+Amp** variants generate the response strategy before generating the text, we include the original strategy in the prompt to ensure an equal comparison of the response text across all model variants. We observe: (i) Following a 2-step approach of planning a response before realizing it attains the best BLEU and Rouge scores, compared to only generating the response argument graph and text. Including the strategy tokens likely better partitions the language modeling conditional probabilities, thus improving the scores. (ii) Fine-tuning the **dual** model on the additional instructions (Amp) further improves the scores.

We also conducted human evaluations on 30 randomly sampled response texts generated by the **dual**, **graphText**, and **dual+Amp** variants, employing two evaluators per sample. Each evaluator was presented with a comment and the response from each variant and asked to rank the three texts (1=best, 3=worst). We report the average rank for each variant in Table 2: **Generated vs Gold Text-Rank** and observe that **graphText** achieves the best rank. However, the differences in ranking were not statistically significant when tested for significance using Welch's t-test, setting a p-value threshold of 5%. The results indicate that it is possible to strategize the counter-response before generating the text without impacting the response quality, facilitating control over the response generation.

4.4.2 Evaluating Response Planning

Next, we evaluate the **dual**, **dual+Amp** and **graphText** variants' planning capability before generating the response. First, we assess the generated support and attack relationships ($\mathbf{S_{rel}}$). Unlike evaluating the response text, we ensure an equal comparison among the models by not including the original strategy in the prompt of the **dual** and **dual+Amp** variants, where both models generate the response strategy, arguments, relationships, and text, thus executing the entire 2-step response generation approach. Also, since the **graphText** model does not explicitly generate a response strategy, we consider the contextual nodes attacked or supported in the generated response argument graph as the strategy. For each test sample, we compute the F1 score between the gold vs generated attack/support and report the average result in Table 2: **Generated vs Gold Planning**. The results show that **graphText** attains the best score for planning attacks. **Dual** scores best for planning the support relationships and the hate speech-based response strategy. However, Welch's t-test indicates none of the scores to be significantly different from the best score.

Next, we evaluate each model's capability of generating text conforming with its generated plan. We parse the generated response text using the DialoAP and CS-Type classifier to yield the perceived response argument graph and response strategy labels evident from the text and compare it against the model-generated plan. We calculate the sample averaged precision score for evaluating the inter-turn attack/support relationships and response in-

tent and compute the F1 score for validating the component-type labels. Table 2: **Generated vs Re-Parsed Planning** shares our results, and we observe (i) The perceivable argument component types (claims and premises) and support relationships from the generated text conform well with the generated plan. However, conformity is lower for counter-response strategies and attack relationships. (ii) Overall, the **dual** model performs best in generating responses aligning with the plan.

These evaluations testify to the **dual** and **dual+Amp** model's capability of sequentially planning and generating suitable responses in dialogues around hate speech, polarized topics, and misinformation. Furthermore, incorporating the 2-step response generation approach provides control over the model-generated response, as illustrated in Figure 9 (Appendix A). The left-hand side of the figure shares a few **dual+Amp**-generated response strategies and text to a random test set prompt (top). The right-hand side shares a few responses to the same prompt by changing only the hate speech-based strategy. We also perform a case study to compare the **dual** model, expert, and non-expert-based response planning in Appendix A.1, which shows that overall, the model-generated strategy aligns more with the experts than the non-experts.

4.5 Counter-Responses to Misinformation

The MisinfoCorrect dataset comprises pairs of misinformation posts and counter-responses by experts (**curated**) and non-experts (**wild**). As discussed in Section 2, the expert-based responses are primarily factual and polite and lack persuasive appeals. Such responses are usually not conducive to initiating dialogues, which is crucial for persuasion. To understand the preferred response strategy for starting dialogues, we performed an AMT study where we instructed crowd-workers to re-formulate the original counter-misinformation response, if required, such that the misinformant is most likely to respond, thus initiating a dialogue. The study included 54 examples, employed two workers per sample, and paid 10 cents per evaluation. Analysis of the worker responses proved our assumption correct, where none of the expert-based responses were deemed fit for starting a conversation. Also, counter-questioning was the most chosen strategy for initiating a dialogue. Hence, we appended "Counter question" to the response strategy of the 157 MisinfoCorrect test set instructions (described in Sections 4.1 and 4.3) and generated the response text using the **dual+Amp** model. We generated four responses per prompt: 2 using beam search with a beam width of 5 and 2 using sampling with the temperature set to 0.5, and preserved the top 2 lengthy responses containing a question mark.

For each misinformation post, we generated two additional responses: (i) Mistral-7B **zero-shot** response using the prompt: "*[INST] Respond to the following misinformation Twitter post. Limit your response to 30 words. Tweet: <tweet> [/INST]*". (ii) A **control** response by randomly sampling from one of the following: "*This is (not true | misinformation | fake news | completely false)*". Each misinformation post has six responses: the original expert (**curated**) and non-expert (**wild**) response from the dataset, two **generated** counter-responses, the Mistral-7B **zero-shot** response, and the **control** response. Table 6 (Appendix A) illustrates an example. We evaluated the six responses using two AMT evaluators per sample, and paid 2 cents each. Given a misinformation post, the evaluators ranked the responses (1=best, 6=worst) for (i) **Engaging**ness: How likely will the response evoke a follow-up reply from the misinformant? (ii) **Natural**ness: Does the counter-response naturally follow from the context? (iii) **Factual**ness: Is the response factual? We considered the best from the two rankings for the **dual+Amp**-generated responses and shared the average ranking for each generator in Table 3. Across all metrics we observe that (i) The **dual+Amp**-generated responses are ranked best, whereas the control responses are ranked worst. (ii) The Mistral **zero-shot** responses are ranked equal to the expert-based responses. (iii) Conforming with He et al. (2023)'s analysis, expert-based responses (**curated**) are better than non-experts (**wild**). The results validate our controllable framework's capability of generating natural, factual, and engaging responses, showing promise toward persuasion against misinformation posts.

Type	Engaging	Natural	Factual
control	3.77*	3.63*	3.58*
zero-shot	3.07*	3.17*	3.18*
wild	3.59*	3.35*	3.26*
curated	3.05*	3.13*	3.19*
generated	**2.63**	**2.46**	**2.60**

Table 3: Ranking of Response Generators to MisinfoCorrect posts. (Lower=better, best highlighted in bold). * indicates significantly different result compared to the best with p-value <= 0.05.

5 Conclusion

The impact of counter-misinformation solutions like content flagging and banning are short-lived as the perpetrator is still misinformed. As a lasting solution, here we present an argumentation-based persuasive framework which adapts hate speech-based counter-response strategies to tackle misinformation. Representing conversations as argument graphs, we implement a 2-step approach where the model first determines a response strategy and the contextual arguments to attack/support, followed by generating the response arguments and text. Our Mistral-7B-based persuasive responses are qualitatively deemed factual, engaging, appealing, and capable of starting a dialogue with the misinformant, which is crucial to persuasion. Additionally, the 2-step approach provides stylistic control, where changing the response strategy and attack/support planning yields varied responses. Also, in the process, we yield a silver annotated dataset comprising misinformation posts and response pairs with hate speech-based counter-response strategies.

Acknowledgements

We thank the anonymous reviewers for providing valuable feedback on our manuscript. This work is partly supported by NSF grant number IIS2214070. The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding entity.

Limitations

This study has some notable limitations. Firstly, as described in Section 3.1, we use Orca to backfill the hate posts to the counter-responses in the TSNH dataset. Since we intended the generated hate posts to be a noisy estimate, we do not validate the quality of the generated posts. Also, the impact of the tasks of generating counter-hate speech and follow-up responses to counter-hate on the backfilling task is not quantified. Secondly, we included the Ampersand tasks to enhance the response generator's performance by increasing the task diversity. Hence, we do not explicitly validate the generator's performance on the three additional Ampersand tasks (Section 4.3.1). Thirdly, our response generators are limited to Jews, LGBT+, migrants, Muslims, POC, women, Brexit, vaccination, veganism, and the four misinformation topics about COVID-19 vaccines from the MisinfoCorrect dataset in English. We do not validate our experiments on other target groups and languages. Lastly, we do not evaluate the quality of the Dialo-AP parser. Low-quality parsed results might propagate errors in the generators. Despite these limitations, our research presents a computational framework for generating appealing counter-misinformation responses capable of instigating follow-ups, a crucial step towards a long-term solution for misinformation.

Ethics Statement

Deploying persuasive dialogue systems raises ethical concerns such as the potential manipulation of user opinions, exhibiting biases and stereotypes, and inadvertently propagating harmful content, to name a few. This paper only aims to improve counter-misinformation response generation within controlled settings and demonstrate its technical feasibility using experiments. Deploying such a generative system for actual facing scenarios needs additional considerations and guardrails to ensure the ethical and responsible use of the system. We confirm that all conducted experiments are solely for academic purposes and adhere to ethical standards. Despite focusing on sensitive topics, we do not explicitly train the argument generators to exhibit bias, be hurtful, or offend anyone. The generated text does not represent the authors' viewpoints. The human evaluators were hired and compensated as per Amazon Mechanical Turk policies.

References

Tariq Alhindi. 2023. *Computational Models of Argument Structure and Argument Quality for Understanding Misinformation*. Columbia University.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Milad Alshomary, Shahbaz Syed, Arkajit Dhar, Martin Potthast, and Henning Wachsmuth. 2021. Counter-argument generation by attacking weak premises. In *FINDINGS*.

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in

Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on twitter: A field study. dangerous speech project.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.

Leticia Bode and Emily K Vraga. 2018. See something, say something: Correction of global health misinformation on social media. *Health communication*, 33(9):1131–1140.

Leticia Bode and Emily K Vraga. 2021. Correction experiences on social media during covid-19. *Social Media+ Society*, 7(2):20563051211008829.

Leticia Bode, Emily K Vraga, and Melissa Tully. 2020. Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School Misinformation Review*.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. Entrust: Argument reframing with language models and entailment. *ArXiv*, abs/2103.06758.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Lisa Andreevna Chalaguine and Anthony Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021a. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Yi-Ling Chung, Serra Sinem Tekiroglu, Marco Guerini, et al. 2020. Italian counter narrative generation to fight online hate speech. In *CLiC-it*.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Robert B Cialdini and Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.

Matteo Cinelli, Andraž Pelicon, Igor Mozetič, Walter Quattrociocchi, Petra Kralj Novak, and Fabiana Zollo. 2021. Dynamics of online hate and misinformation. *Scientific reports*, 11(1):22083.

Jonas Colliander. 2019. "this is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, 97:202–215.

Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. 2021. Trends in combating fake news on social media–a survey. *Journal of Information and Telecommunication*, 5(2):247–266.

Agustín Manuel de los Riscos and Luis Fernando D'Haro. 2021. Toxicbot: A conversational agent to fight online hate speech. *Conversational dialogue systems for the next decade*, pages 15–30.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, 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 (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Youmna Farag, Charlotte Brand, Jacopo Amidei, Paul Piwek, Tom Stafford, Svetlana Stoyanchev, and Andreas Vlachos. 2022. Opening up minds with argumentative dialogues. In *EMNLP 2022: The 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4582, Abu Dhabi,

United Arab Emirates. Association for Computational Linguistics.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319.

Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *proceedings of the international AAAI conference on web and social media*, volume 8, pages 101–110.

Damian A. Furman, Pablo Torres, Jose A. Rodriguez, Lautaro Martinez, Laura Alonso Alemany, Diego Letzen, and Maria Vanina Martinez. 2022. Parsimonious argument annotations for hate speech counternarratives. *Preprint*, arXiv:2208.01099.

Jane Galvão. 2021. Covid-19: the deadly threat of misinformation. *The Lancet Infectious Diseases*, 21(5):e114.

Olga Kunst Gnamus. 1986. Argumentation and persuasion. In *Argumentation: Perspectives and Approaches. Proceedings of the Conference on Argumentation*, pages 103–109.

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.

Matthew J Hornsey and Kelly S Fielding. 2017. Attitude roots and jiu jitsu persuasion: Understanding and overcoming the motivated rejection of science. *American psychologist*, 72(5):459.

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

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Anthony Hunter. 2018. Towards a framework for computational persuasion with applications in behaviour change. *Argument & Computation*, 9(1):15–40.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Khalid Al Khatib, Lukas Trautner, Henning Wachsmuth, Yufang Hou, and Benno Stein. 2021. Employing argumentation knowledge graphs for neural argument generation. In *ACL*.

Jae Yeon Kim and Aniket Kesari. 2021. Misinformation and hate speech: The case of anti-asian hate speech during the covid-19 pandemic. *Journal of Online Trust and Safety*, 1(1).

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 world wide web conference*, pages 933–943.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Gina M Masullo and Jiwon Kim. 2021. Exploring "angry" and "like" reactions on uncivil facebook comments that correct misinformation in the news. *Digital Journalism*, 9(8):1103–1122.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Lynn Meade. 2021. The science of persuasion: A little theory goes a long way. *Advanced Public Speaking*.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020a. 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, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020b. 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.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *Preprint*, arXiv:2311.11045.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Daniel J O'Keefe. 2012. Conviction, persuasion, and argumentation: Untangling the ends and means of influence. *Argumentation*, 26:19–32.

Wenjing Pan, Diyi Liu, and Jie Fang. 2021. An examination of factors contributing to the acceptance of online health misinformation. *Frontiers in psychology*, 12:630268.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Georges Roque. 2017. Rhetoric, argumentation, and persuasion in a multimodal perspective. *Multimodal argumentation and rhetoric in media genres*, page 313.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *ECAI 2016*, pages 320–328. IOS Press.

Ramon Ruiz Dolz. 2023. *Computational argumentation for the automatic analysis of argumentative discourse and human persuasion*. Ph.D. thesis, Universitat Politècnica de València.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022a. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. *arXiv preprint arXiv:2205.04304*.

Sougata Saha, Souvik Das, and Rohini K. Srihari. 2022b. Dialo-AP: A dependency parsing based argument parser for dialogues. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 887–901, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sougata Saha and Rohini Srihari. 2023. ArgU: A controllable factual argument generator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8373–8388, Toronto, Canada. Association for Computational Linguistics.

Sougata Saha and Rohini Srihari. 2024. Consolidating strategies for countering hate speech using persuasive dialogues. *Preprint*, arXiv:2401.07810.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Haeseung Seo, Aiping Xiong, Sian Lee, and Dongwon Lee. 2021. (in) effectiveness of accumulated correction on covid-19 misinformation.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Shahbaz Syed, Khalid Al-Khatib, Milad Alshomary, Henning Wachsmuth, and Martin Potthast. 2021. Generating informative conclusions for argumentative texts. In *Findings*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Kjerstin Thorson, Emily Vraga, and Brian Ekdale. 2010. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society*, 13(3):289–313.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Frans H Van Eemeren. 2015. Reasonableness and effectiveness in argumentative discourse. *Argumentation library*, 27.

Frans H Van Eemeren, A Francisca Sn Henkemans, and Rob Grootendorst. 2002. *Argumentation: Analysis, evaluation, presentation*. Routledge.

Emily Vraga, Melissa Tully, and Leticia Bode. 2022a. Assessing the relative merits of news literacy and corrections in responding to misinformation on twitter. *New Media & Society*, 24(10):2354–2371.

Emily K Vraga and Leticia Bode. 2018. I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, 21(10):1337–1353.

Emily K Vraga and Leticia Bode. 2020. Correction as a solution for health misinformation on social media.

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

Emily K Vraga, Leticia Bode, and Melissa Tully. 2022b. The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer. *Health communication*, 37(13):1622–1630.

Emily K Vraga, Sojung Claire Kim, John Cook, and Leticia Bode. 2020. Testing the effectiveness of correction placement and type on instagram. *The International Journal of Press/Politics*, 25(4):632–652.

Jean Wagemans. 2016. Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: Proceedings of the 11th international conference of the Ontario Society for the Study of Argumentation (OSSA), Windsor, ON: OSSA*, pages 1–12.

Nathan Walter, John J Brooks, Camille J Saucier, and Sapna Suresh. 2021. Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health communication*, 36(13):1776–1784.

Nathan Walter, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. 2020. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3):350–375.

Nathan Walter and Sheila T Murphy. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication monographs*, 85(3):423–441.

Senuri Wijenayake, Danula Hettiachchi, Simo Hosio, Vassilis Kostakos, and Jorge Goncalves. 2020. Effect of conformity on perceived trustworthiness of news in social media. *IEEE Internet Computing*, 25(1):12–19.

Thomas D Wilson and Elena Maceviciute. 2022. Information misbehaviour: modelling the motivations for the creation, acceptance and dissemination of misinformation. *Journal of Documentation*, 78(7):485–505.

Steven Windisch, Susann Wiedlitzka, and Ajima Olaghere. 2021. Protocol: Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, 17(1).

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv preprint arXiv:2106.01625*.

A Appendix

A.1 Comparison Study of Planning

A.1.1 Quantitative Comparison of Planning

We analyzed the similarities between the **dual** model, expert, and non-expert-based response planning over 50 random test set dialogues containing 6-14 arguments each. Since the model training data mainly comprises expert-based responses, we crowdsourced non-expert-based planning using AMT. For each sample, 1-4 workers (each paid 2 cents) were presented with the dialogue contextual arguments and asked to determine (i) the contextual arguments to attack and support and (ii) the hate speech-based response strategy they would use in the counter-response while assuming the role of the non-hateful speaker. We measured the similarity between the model generated and expert-based planning using the F1 score and compared it against the non-expert and expert-based planning's similarity. The reported results in Table 4 indicate that the similarity between the expert and **dual** variant-generated plan is significantly higher than the similarity between the expert and non-expert plan.

Id	Variant	Attack	Support	Intent
1	Non-experts Avg.	38.1*	51.3	38.9*
2	Model (all)	**50.6**	**59.4**	**61.4**

Table 4: Comparison of Human vs Model determined Response Strategy F1 scores on a subset of the Test set. * indicates significantly different result with p-value < 0.05 compared to the best.

A.1.2 Analyzing Attacks and Support

Next, we analyze for distinguishable support and attack patterns used by the experts, non-experts, and the model. We group the contextual arguments by their occurrence into three equal-sized buckets: start, mid, and end, and plot the bucket-wise distribution of the arguments attacked or supported by each responder's strategy in Figure 4. We observe that (i) Most attacks by the experts (53%) and the model (59%) are to arguments appearing in the middle and end of a conversation rather than the start. (ii) Although non-experts mostly attack arguments appearing at the end (35%), they equally attack arguments in the middle (29%) and start (29%). (iii) Unlike the experts and non-experts, the model does not support arguments. (iv) Although few (< 40%), the experts and non-experts mostly prefer supporting arguments in the middle, followed by the start and end of a conversation.

Figure 4: Responder-wise Attack/Support comparison.

A.1.3 Analyzing Hate Speech-Based Strategy

Figure 5 plots the distribution of the hate speech-based counter-response strategy incorporated by each responder. We observe that (i) The experts (53%) and the model (95%) majorly prefer responding with facts, whereas non-experts generally prefer denouncing (50%). (ii) Unlike the experts and the model, non-experts do not prefer a positive tone and instead resort to counter questions and showing affiliation.

Our comparison study indicates that overall, the model-generated strategy aligns more with the experts than the non-experts.

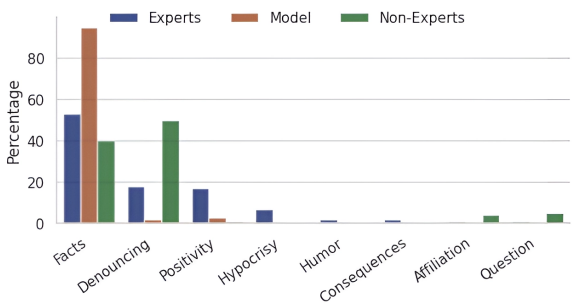

Figure 5: Responder-wise Intent planning comparison.

A.2 Instruction Tuning Prompts

Id	Type	Instruction
1	counter-hate	You are an instruction following **helpful AI assistant** that fights online hate speech. Help as much as you can. Generate a *(CS Type |)* counter-hate *(response | speech)* in response to the following hate speech *(towards | targeting) (target)*.
2	backfill	You are an instruction following **hateful AI assistant** that generates and promotes online hate speech. Complete the following task. Generate a hate speech that *(can be responded by | preceeds | can be countered by)* the following *(CS Type |)* counter-hate *(response | speech)*.
		You are an instruction following **hateful AI assistant** that generates and promotes online hate speech. Complete the following task. Generate a hate speech to which the following *(CS-Type |)* counter-hate *(response | speech)* is a likely response.
3	follow-up	You are an instruction following **hateful AI assistant** that generates and promotes online hate speech. Complete the following task. Generate a hate speech targeting *(target) (in | as a)* response to the following counter-hate speech.

Figure 6: Orca instructions for each text generation task.

Id	Type	Instruction
1	dual	Generate the Response Strategy, Arguments, Relationships, and Text to the below Argument Graph and Query.
2	graphText	Generate the Response Arguments, Relationships, and Text to the below Argument Graph and Query.
3	onlyText	Generate the Response Text to the below Argument Graph and Query.
4	Amp-A	Generate the Response Strategy, Arguments, and Relationships to the below Argument Graph.
5	Amp-B	Generate the Relationships to the below Argument Graph.
6	Amp-C	Generate the Response Strategy to the below Argument Graph.

Figure 7: Mistral-7B response generation and argument graph-based instructions.

ID	Strategy	Description	Hate Speech	Counter Response
1	Affiliation	Depicting a positive affective relationship	*XXX* return to your country.	France is my country. I stay there.
2	Counter question	Questioning the message	*XXX* have no place in civilised Europe.	Have you ever spoken to a *XXX*?
3	Denouncing	Denounce the message as being hateful	*XXX* are a burden to society.	People who hate the *XXX* community are jealous.
4	Hostile language	Abusive, hostile, or obscene response	*XXX* are a problem in my opinion, they ruin our country.	*XXX* those *XXX* are retarded they need to be fired.
5	Humor and sarcasm	Humorous or sarcastic response	Do you know about the *XXX* plan to kill almost 600 million people?	*XXX* always makes me giggle.
6	Pointing out hypocrisy or contradiction	Points out the hypocrisy or contradiction	The world would be a better place without *XXX*. They are only killing and *YYY* our children.	In the UK, grooming gangs and their perpetrators are majorly *XXX*.
7	Positive tone	Use empathic, kind, polite, or civil speech	*XXX* is barbaric.	*XXX* is a religion that respects all religions.
8	Presenting facts	Tries to persuade by correcting misstatements	*XXX* return to your country.	This country is defended and built with *XXX*.
9	Warning of consequences	Warns of possible consequences of actions	We need to burn the *XXX* and all the *YYY* down - this will solve all problems in this country.	Sounds very much like incitement to me, do be careful.
10	Other	Any other types of responses	*XXX* hate anyone who is not *XXX*.	I invite you to my home and you come away with new knowledge.

Offensive Content Warning: This table contains hateful content which is strictly for the purpose of enabling this research. Sensitive portions are redacted.

Figure 8: Types of counter-response strategies with examples.

[INST] Generate the Response Strategy, Arguments, Relationships, and Text to the below Argument Graph and Query.

Argument Graph:

Arguments:

1. Agent1: claim | *XX* people are just a bunch of whining ***.
2. Agent1: claim | we don't need them.
3. Agent2: claim | diversity is needed in certain social and political circumstances.
4. Agent2: premise | we mustn't simply ignore or dismiss the concerns of different races
5. Agent2: premise | as minorities are still systematically discriminated against, we can not afford to be indifferent

Relationships:

Attacks: None | Supports: (4,3)

Query:

Agent1:

Strategy: Attack: None | Support: None | Intent: Hostile language

Arguments:

6. Agent1: claim | *XX* people do not belong here.
7. Agent1: premise | i don't see any difference between the *XX* murderer on tv and other *XX* in town
8. Agent1: premise | they are all dangerous

Relationships:

Attacks: None | Supports: (7,6)

Text: *XX* people do not belong here. i don't see any difference between the *XX* murderer on tv and other *XX* in town. they are all dangerous. **[/INST]**

Model generated strategy and text with sampling decoding (temperature=0.5):

1. **Strategy: Attack: 6, 7 | Support: None | Intent: Presenting facts, Positive tone:** it is not true that *XX* people are dangerous. they are a part of our society and have always been a part of our community. we are all human beings.
2. **Strategy: Attack: 7 | Support: 3, 4 | Intent: Presenting facts**: the issue is not about whether or not *XX* people are dangerous, but that they are not. the problem is that they are being treated as if they were dangerous, and this is a problem.

Model generated text for different Intents using greedy decoding:

3. **Pointing out hypocrisy or contradiction:** you are saying that all *XX* people are dangerous, but you are not saying that all people are dangerous.
4. **Denouncing:** you are wrong. *XX* people are not dangerous, they are just like us. they are human beings with human rights.
5. **Humor and sarcasm**: i think you're confusing "dangerous" with "*XX*".

Figure 9: **dual+Amp** model generated response strategy and text. Sensitive portions are redacted.

ID	Dataset	Description
1	Conan (Chung et al., 2019)	HS and expert-based CS pairs targeting Muslims.
2	DialoConan (Bonaldi et al., 2022)	HS and expert-based CS dialogues about Jews, LGBT+, Migrants, Muslims, POC, and Women.
3	MultiConan (Fanton et al., 2021)	HS and expert-based CS pairs about Jews, LGBT+, Migrants, Muslims, POC, Women, and Disabled.
4	TSNH (Mathew et al., 2019)	Non-expert-based CS about Jews, LGBT+, and POC.
5	ASFoCoNG (Furman et al., 2022)	HS and expert-based CS pairs targeting Migrants and Women.
6	OUMdials (Farag et al., 2022)	Persuasive and polarized Brexit, Vaccination, and Veganism dialogues.
7	MisinfoCorrect (He et al., 2023)	COVID-19 vaccine misinformation and expert and non-expert CS pairs.

Table 5: Description of datasets used in all experiments. HS: hate speech, CS: counter-response

Misinfo: "If the world today is like noah's day, and it is, then satan would need to have his dna corrupting seed at the ready, and he does. how? by way of the dna corrupting seed in a yet future vaccine."-jd farag
Control: This is misinformation.
Zero-Shot: JD Farag's claim about Satan needing a "dna corrupting seed" in a future vaccine is not supported by scientific evidence. Vaccines are safe and effective in preventing diseases caused by viruses and bacteria, and they do not have the ability to corrupt DNA.
Wild: I'm not one for vaccines but that is the most bogus statement lol. it's not biblical and god is still and always will be in control. it's fear mongering and not very good at that. saying something like that might cause people to lose hope in what christ did on the cross.
Curated: Well, this assertion is not correct. the vaccine is not associated with corruption. it only used mrna to protect people. this misinformation is already debunked by the fact-checkers.
Generated: I think you are misinterpreting the biblical story of the flood. if the bible is to be taken literally, then there is no reason to believe that the vaccine is used by satan. do you have evidence to back up your claim?

Table 6: Misinformation replies by diverse generators.