# Repairing Fallacious Argumentation in Political Debates

Pierpaolo Goffredo, Deborah Dore, Elena Cabrio, Serena Villata

# Repairing Fallacious Argumentation in Political Debates

Pierpaolo Goffredo
*Université Côte d'Azur, CNRS, Inria, I3S, France*
pierpaolo.goffredo@univ-cotedazur.fr

Deborah Dore
*Université Côte d'Azur, CNRS, Inria, I3S, France*
deborah.dore@univ-cotedazur.fr

Elena Cabrio
*Université Côte d'Azur, CNRS, Inria, I3S, France*
elena.cabrio@univ-cotedazur.fr

Serena Villata
*Université Côte d'Azur, CNRS, Inria, I3S, France*
serena.villata@cnrs.fr

## Abstract

Fallacies, defined as arguments based on erroneous logical foundations, are pervasive in political discourse due to their persuasive nature and illusion of validity. Such misleading content and its spread may seriously influence and impact societal well-being by leading to inaccurate conclusions and invalid inferences from citizens and policymakers. Automatically detect and defuse fallacious arguments is therefore crucial to limit the spread of manipulative claims and to promote a healthier political debate. In this paper, we address such challenge, casting it as the computational task of *repairing* fallacious arguments in political debates. Our contribution in this novel task is manifold. Firstly, we introduce a new dataset, **FallacyFix**: A Repaired Fallacies Dataset, comprising repaired examples across various fallacy categories. Secondly, we propose a series of prompt techniques for generating non-fallacious arguments, (in)dependent of the fallacy label being addressed. Third, we introduce a novel evaluation methodology to assess the quality of the generated text, specifically designed to repair fallacies in political debates. Lastly, we perform a user study to assess the Relevance, Suitability, and Cogency of the generated repaired arguments.

# 1 Introduction

Fallacious arguments [Oswald and Herman, 2020] are defined as "invalid" arguments (e.g., the conclusion does not follow from the premises) [Hamblin, 1970] or wrong moves in argumentative discourse [Eemeren and Grootendorst, 1987]. This kind of argumentation is therefore misleading or deceptive, in particular when employed in political debates. Although some scholars have casted doubt on the ubiquity and deceptiveness of fallacies [Woods, 2013, Boudry et al., 2015] and believe the resulting concerns to be overblown [Paglieri, 2019], the prevalent opinion is that spreading of this nefarious content severely impacts the society and the decision-making of both citizens and policymakers: on this bleaker view of fallacies, it is vital to prevent fallacious and propagandist arguments to circulate. To address this challenging task, several approaches proposing to identify fallacious argumentation in text have been presented in the literature [Ruiz-Dolz and Lawrence, 2023, Chen et al., 2024, Li et al., 2024a, Helwe et al., 2024, Payandeh et al., 2024, Li et al., 2024b, Goffredo et al., 2023, 2022].

However, merely identifying this content is insufficient to ensure that the audience realizes the impact of the fallacious argument on its deliberation process and to support the development of critical thinking skills Walton [2008]. To tackle this challenging goal [Lewandowsky et al., 2012, Bode and Vraga, 2015, Visser et al., 2020], it is necessary to unveil why a particular argument is fallacious and to demonstrate how it could be repaired as a valid, non-fallacious argument.

In this paper, we address this key challenge by proposing a new task called *repairing fallacious argumentation*. The goal of this task is to modify statements that contain fallacious arguments into versions that are *clearer*, *fairer*, and *free from any technique that could negatively persuade listeners*. As a case study, we carry out this task on political debates, where the need for this kind of solution is urgent. Our contribution in addressing this task is manifold: *i)* a novel dataset, **FallacyFix**, comprising repaired examples across various fallacy categories (Appeal to Fear, Appeal to Pity, Appeal to Popular Opinion, Flag Waving, and Loaded Language) based on the ElecDeb60to20-fallacy dataset [Goffredo et al., 2023];
*ii)* modular prompt techniques for generating non-fallacious arguments, both dependent and independent of the specific fallacy label being addressed. Through an extensive evaluation, we assess these techniques using the most widely used Large Language Models (in Zero-Shot, Few-Shot, and Fine-Tuning settings) and a standard baseline model (BART);
*iii)* a rigorous evaluation framework to assess the accuracy of the generated non-fallacious argument repairing the fallacy in the original argument, with respect to the manually annotated benchmark of non-fallacious arguments we built from the

ElecDeb60to20 dataset (as described in Section 3.2);
*iv)* a human evaluation of the generated non-fallacious arguments to assess the acceptability of these arguments across three dimensions, i.e., Relevance, Suitability, and Cogency.

This paper marks the first attempt to systematically analyze and generate non-fallacious arguments from the identified fallacious ones in political debates, towards a more reliable political discourse.

The paper is organized as follows: Section 2 discusses the works closely related to ours, Section 3 presents the dataset, Section 4 introduces the task of fallacious argument repair, Section 5 addresses our research questions and reports the evaluation and Section 6 concludes the paper and outlines directions for future work.

## 2 Related Works

In recent times, the domain of natural language processing has demonstrated a growing interest in detecting fallacies and associated phenomena, including misinformation and propaganda [Da San Martino et al., 2020]. The pioneering research conducted by Da San Martino et al. [2019] on fallacies present in newspaper articles has served as a significant foundation for motivation within this field. Recently, scholars have achieved remarkable advancements in recognizing and categorizing fallacies embedded within discourse.

Ruiz-Dolz and Lawrence [2023] examined the use of LLMs for detecting argumentative fallacies, testing SVM, fine-tuned RoBERTa [Liu et al., 2019], and Zero-Shot prompting with GPT-3.5 and GPT-4. They found that while models like fine-tuned RoBERTa and GPT-4 performed well on detection datasets, they struggled with deeper logical understanding. Chen et al. [2024] investigated LLMs in computational argumentation tasks, introducing "counter speech generation" to evaluate argument mining (AM) and argument generation (AG). LLMs outperformed baselines on some datasets, achieving high BERTScore, but faced challenges with strict metrics like ROUGE. Li et al. [2024a] proposed tasks from cognitive dimensions to enhance logical fallacy understanding. They fine-tuned LLMs with targeted datasets, which consistently improved performance in logical reasoning tasks compared to using original data alone. Helwe et al. [2024] presented a unified benchmark for fallacy detection and classification, including a new annotated dataset and evaluation of modern LLMs and humans. They found that Zero-Shot detection was somewhat feasible, but fine-grained classification remained difficult for both LLMs and humans. Payandeh et al. [2024] evaluated the reasoning capabilities and susceptibility to logical fallacies of GPT-3.5 and GPT-4 using the LOGICOM benchmark.

Their findings revealed that LLMs are more susceptible to fallacious arguments than to logical reasoning, with GPT-4 being more susceptible than GPT-3.5. Li et al. [2024b] introduced the FLUB dataset, designed to test LLMs' understanding of fallacies using cunning questions from an online forum. Their evaluation showed that current models struggle with fallacies and cunning questions, highlighting the need for improved reasoning capabilities in LLMs. Goffredo et al. [2023] focused on detecting and classifying fallacious arguments in U.S. presidential debates from 1960 to 2020, including the 2020 Trump-Biden debates. They developed MultiFusion BERT, a transformer-based model that integrates text representations with argumentative features and PoS tags, enhancing fallacy detection and classification. Ramponi et al. [2025] introduce FAINA, a dataset for fallacy detection that embraces multiple plausible answers and natural disagreement.

Despite significant progress in fallacy detection and reasoning, the field has primarily concentrated on identifying and classifying fallacies rather than correcting them. We address this gap by introducing an approach to automatically repair fallacious arguments, preserving their core claims while eliminating logical flaws.

## 3 FallacyFix: A Repaired Fallacies Dataset

To address the task of repairing fallacious arguments within political debates, we relied on the ElecDeb60to20-fallacy dataset [Goffredo et al., 2023]. This dataset comprises televised debates from U.S. presidential election campaigns spanning from 1960 to 2020. The debates were sourced from the website of the Commission on Presidential Debates[1], which provides publicly accessible transcripts of debates featuring prominent candidates for presidential and vice-presidential nominations in the United States. This temporal span ensures a comprehensive representation of political rhetoric, making it a valuable resource for analyzing fallacies in political discourse.

The ElecDeb60to20-fallacy dataset[2] includes the following main categories of fallacies: *Appeal to Emotion, Appeal to Authority, Ad Hominem, False Cause, Slippery Slope*, and *Slogans*. These labels are based on 2,744 annotated arguments, providing a robust framework for identifying and analyzing different types of fallacious reasoning within the debates.

[1] `https://www.debates.org/`

[2] `https://github.com/pierpaologoffredo/ElecDeb60to20`

### 3.1 Fallacious Argument Annotation

We relied on the fallacies annotated in the ElecDeb60to20-fallacy dataset as the source to generate *repaired* fallacious arguments, that we collect in a new dataset named **FallacyFix**: A Repaired Fallacies Dataset[3]. Given that the ElecDeb60to20-fallacy dataset contains multiple categories of fallacy with different levels of complexity, we first conducted a pilot study to evaluate the feasibility of the task on a few of them, also in terms of human annotation effort and time efficiency.

For instance, if we take the following the fallacious *Flag Waving* argument:
*"That's dangerous, and it's provocative. And the mixed message, the ambiguities of U.S. foreign policy, are –* ***I believe, and Bob Dole believes, is causing not only problems for this country throughout the world, but particularly here at home.*** *And the type of changes that were made overnight in California caused very severe dislocations."*[4]
The repaired version could be written as follows:
*"That's dangerous and it's provocative. And the mixed message, the ambiguities of U.S. foreign policy, are –* ***This is causing problems everywhere.*** *And the type of changes that were made overnight in California caused very severe dislocations."*

To start with, we considered *Appeal to Emotion* and *Appeal to Authority* as the main categories to focus on. From these, we selected specific related subcategories for our study: *Loaded Language, Flag Waving, Appeal to Pity, Appeal to Fear* (subcategories of Appeal to Emotion), and *Appeal to Popular Opinion*, *Without Evidence*, and *False Authority* (subcategories of Appeal to Authority). Through this preliminary study, we identified the most relevant and feasible subcategories to include in our final annotation guidelines. We decided to focus on these subcategories to explore the task of fallacy repairing because *(1)* research indicates that emotional appeals and authoritative references are more straightforward for annotators to identify and categorize due to their distinct and recognizable features [Habernal et al., 2018a], and *(2)* other categories, such as Ad Hominem and False Cause, necessitate a more nuanced contextual understanding, which would substantially increase the complexity and effort required for annotation [Eemeren and Grootendorst, 2003, Stab and Gurevych, 2016, Habernal et al., 2018b]. Thus, given the complexity of this novel task, we decided to focus on the (sub)categories on which human agreement was higher.

Below, we present a brief description of each of the final annotated subcategories, along with an illustrative example. In each example, the fallacious segment is shown in bold, and the surrounding context is included for clarity.

[3] `https://github.com/pierpaologoffredo/repairing_fallacies`
[4] Jack Kemp, 09/10/1996.

- **Appeal to Fear:** This fallacy seeks to build support for an idea by instilling anxiety or panic in the population regarding an alternative.

  Example: *"Franklin Roosevelt said in 1932 that the only thing we have to fear is fear itself.* ***The only thing, Jim, they have to offer is fear.*** *Fear of the environment, fear of the climate, fear of Medicare, fear of Newt, fear of Republicans, fear of Bob, and probably fear of cutting tax rates."*

- **Appeal to Pity:** This fallacy involves evading relevant considerations needed to make a decision on an issue by appealing to pity.

  Example: *We are seeking to have vigorous enforcement of the laws that bar discrimination. Look, that's where World War I started in the Balkans.* ***My uncle was a victim of poisonous gas there.*** *Millions of Americans saw the results of that conflict.*

- **Appeal to Popular Opinion:** This fallacy attempts to reinforce political claims by referencing the popularity of an idea or the will of the people.

  Example: *I also believe that it is – one of the reasons we can't do it is we're overextended.* ***Ask the people in the armed forces today.*** *We've got Guards and Reserves who are doing double duties.*

- **Flag Waving:** This propaganda technique appeals to a group by using emotionally charged arguments related to nation, race, gender, political preference, or other group affiliations.

  Example: *So whether it's diabetes or cancer, they come to hospitals later and it costs America more.* ***We got to have health care for all Americans.*** *I think it's important, since he talked about the Medicare plan, has he been in the United States Senate for 20 years?*

- **Loaded Language:** This fallacy involves the use of specific words and phrases with strong emotional implications (positive or negative) to influence an audience.

  Example: *You just heard the president say that young people ought to be able to take money out of Social Security and put it in their own accounts.* ***Now, my fellow Americans, that's an invitation to disaster.*** *The CBO said very clearly that if you were to adopt the president's plan, there would be a $2 trillion hole in Social Security, because today's workers pay in to the system for today's retirees.*

### 3.2 Manually repairing fallacious arguments

The goal of the annotation phase is to transform statements that contain fallacious arguments to produce arguments that are *clearer*, *fairer*, and *free from any technique that could negatively persuade listeners*.

Two annotators with background in computational linguistics conducted the annotation process on a subset of the dataset following the methodology defined in the pilot study, to ensure high-quality results. The annotation of fallacious arguments was conducted on statements with the preceding and subsequent sentences relative to the speech turn containing the fallacy. This methodology leverages the context of fallacious arguments to produce repaired arguments. The repair process relies on simple editing operations such as removal, subtraction, and simplification, as shown in Figure 1.

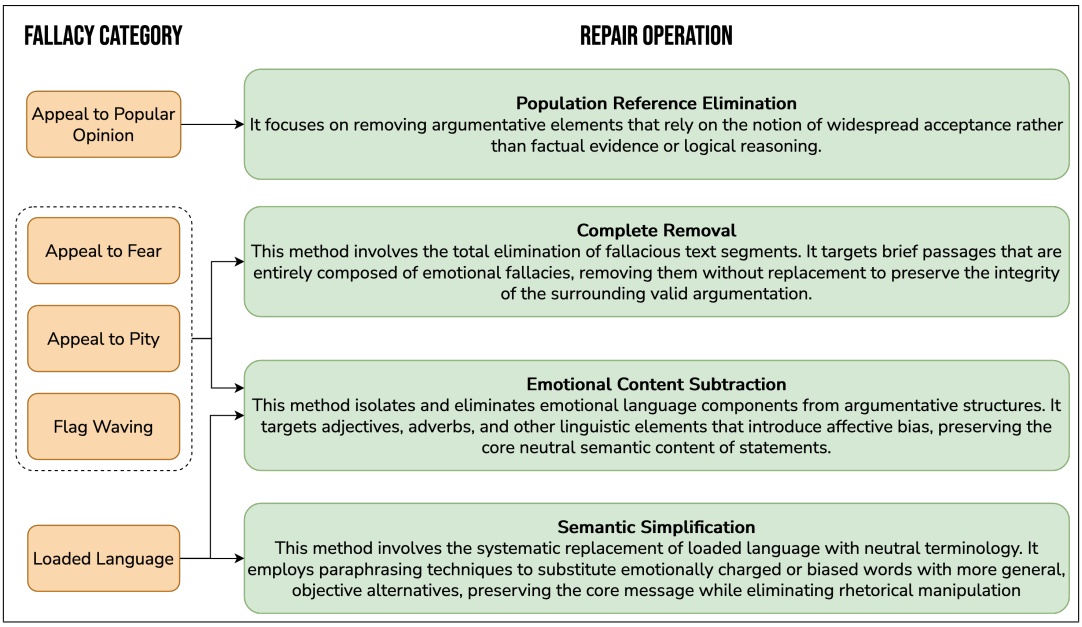


Figure 1: Methodology applied to repair fallacious arguments according to the different category.

According to the proposed schema, *Appeal to Popular Opinion* often makes use of phrases implying widespread acceptance (e.g., "everyone knows"), that is targeted by **Population Reference Elimination**.

Emotional fallacies like Appeals to Fear, Appeal to Pity, and Flag Waving commonly share traits of emotive language: *Appeal to Fear* appeals usually employ

future-tense negative outcomes, *Appeal to Pity* appeals tend to use personal anecdotes and emotive adjectives, and *Flag Waving* generally relies on patriotic jargon. These may undergo **Complete Removal** if entirely emotional, or **Emotional Content Subtraction** if mixed with valid points.

*Loaded Language*, typically marked by biased words and evaluative adjectives, can often be addressed through **Semantic Simplification**, replacing charged terms with neutral alternatives, or **Emotional Content Subtraction** to remove offensive, discriminatory, or inflammatory language while preserving core meaning. Each technique generally aims to preserve the argument structure while removing fallacious elements, guided by the specific syntactic features of each fallacy type.

Both the original fallacious argument and the repaired one have been included into the **FallacyFix** dataset. Figure 2 shows the distribution of the different subcategories for a total of 747 annotated fallacies, each paired with its corresponding repaired version.

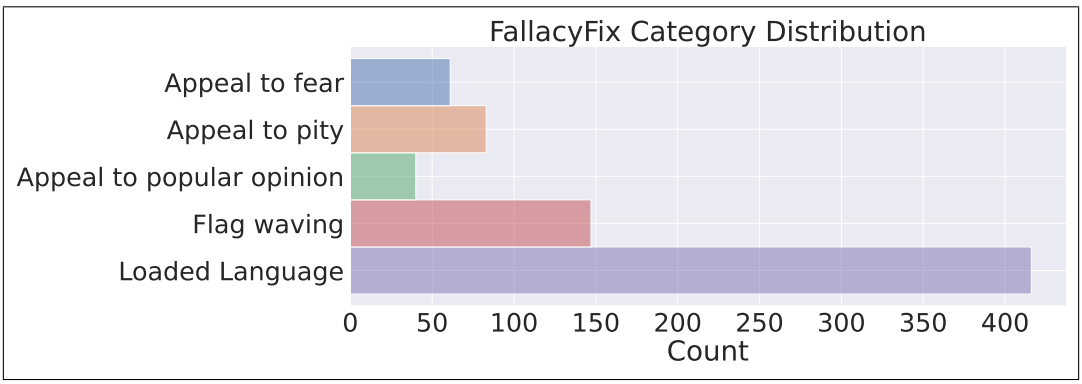


Figure 2: FallacyFix dataset's fallacy category distribution.

This dataset was developed through a two-phase process. Two experts initially annotated a randomly selected set of 100 fallacious arguments from the ElecDeb-60to20-fallacy dataset. These arguments were drawn from the categories of fallacies under analysis. The experts followed the proposed methodology to annotate the arguments, allowing for the establishment of inter-annotator agreement (IAA) between the two resulting repaired texts.[5] To calculate their similarity, we used BERTScore [Zhang et al., 2020] and BERT [Devlin et al., 2019] embeddings, resulting in the following scores: $0.94 \pm 0.06$ and $0.98 \pm 0.03$, respectively.[6] Subsequently, one of the two experts continued the annotation process, expanding it to reach 747

[5]The pair of texts compared rely on the same starting fallacious argument and the same fallacy category to *repair*.

[6]BERTScore was chosen for its ability to capture semantic nuances in context, making it par-

samples. This expert followed the methodologies outlined in the aforementioned schema, applying the appropriate repair techniques for each fallacy category. Table 1 shows examples of repaired fallacies for each category based on the proposed annotation methodology, while Figure 3 complements this by showing how various repair strategies are distributed across the original fallacy subcategories.

| Subcategory | Fallacious Argument | Repaired Argument |
|---|---|---|
| **Appeal to Fear** | The effort that we've mounted with respect to Iraq focused specifically on the possibility that this was the most likely nexus between the terrorists and weapons of mass destruction. **The biggest threat we faced today is the possibility of terrorists smuggling a nuclear weapon or a biological agent into one of our own cities and threatening the lives of hundreds of thousands of Americans.** What we did in Iraq was exactly the right thing to do. | The effort that we've mounted with respect to Iraq focused specifically on the possibility that this was the most likely nexus between the terrorists and weapons of mass destruction. *We faced hard situations.* What we did in Iraq was exactly the right thing to do. |
| **Appeal to Pity** | We just have a different approach. **But let me remind you, my family has suffered from drug abuse. I know what it's like to see somebody you love nearly lose their lives, and I hate drugs, Senator. We need to do this together and we can.** Not if I didn't have a better idea. | We just have a different approach. *We need to do this together and we can.* Not if I didn't have a better idea. |
| **Appeal to Popular Opinion** | No nation will ever have a veto over us. **But I think it makes sense, I think most Americans in their guts know, that we ought to pass a sort of truth standard.** That's how you gain legitimacy with your own countrypeople, and that's how you gain legitimacy in the world. | No nation will ever have a veto over us. *But I think it makes sense, that we ought to pass a sort of truth standard.* That's how you gain legitimacy with your own countrypeople, and that's how you gain legitimacy in the world. |
| **Flag Waving** | That will help reinforce the values that parents teach at home as well. **Ours is a great land, and one of the reasons why is because we're free.** And so I don't support censorship. | That will help reinforce the values that parents teach at home as well. *We live in a democracy.* And so I don't support censorship. |
| **Loaded Language** | And I was so surprised to see him sign on with the devil. But when you talk about apology, I think the one that you should really be apologizing for and the thing that you should be apologizing for are the 33,000 e - mails that you deleted, and that **you acid washed**, and then the two boxes of e - mails and other things last week that were taken from an office and are now missing. And I'll tell you what. | And I was so surprised to see him sign on with the devil. But when you talk about apology, I think the one that you should really be apologizing for and the thing that you should be apologizing for are the 33,000 e - mails that you deleted, and that *you handled in your own way*, and then the two boxes of e - mails and other things last week that were taken from an office and are now missing. And I'll tell you what. |

Table 1: Examples of fallacious arguments and their corresponding repaired versions across different subcategories of fallacies. The text in **bold** represents the original fallacious argument identified by annotators, while the *italic* text indicates the repaired version of the fallacy.

As illustrated in Figure 3, fallacies of the *Loaded Language* type were predominantly repaired through the *Rephrasing strategy.* For *Flag Waving* and *Appeal to*

ticularly suitable for evaluating fallacy repairs where subtle wording changes are crucial. This contextual sensitivity addresses the lack of standardized methods for comparing repaired fallacious arguments.

*Popular Opinion*, annotators most frequently employed *Popular Reference Elimination*. Instances of *Appeal to Pity* were primarily addressed through the *Removal of Additional Information*, whereas *Appeal to Fear* was typically repaired via the *Removal of Emotional Snippets* strategy.

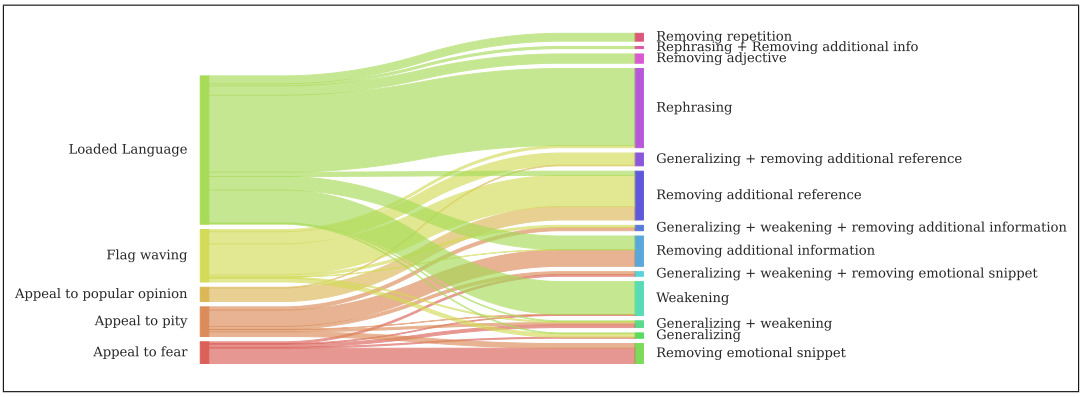


Figure 3: Distribution of the repairing strategies across the original fallacy subcategories.

We split the FallacyFix dataset into training (80%), validation (10%), and test (10%) sets. This split was applied while maintaining a balanced representation of different fallacy categories in each set, ensuring that our models would be trained and evaluated on a diverse range of fallacy types.

## 4 Fallacious arguments repair as a generation task

As introduced before, the task of *repairing* fallacies in political debates involves transforming statements that contain fallacious arguments into versions that are *clearer*, *fairer*, and *free from any techniques that could negatively persuade listeners*. While this task can be highly subjective in principle, in the previous section we have proposed a methodology to ensure that the repaired statements are impartial and devoid of manipulative rhetoric that are present in fallacies [Eemeren, 2001, Eemeren and Grootendorst, 1987]. We approach this as a computational task using large language models (LLMs). These models automatically generate human-like arguments that repair the fallacious nature of given statements. We then compare these generated fallacy-free arguments with the manually created ones in the FallacyFix dataset. Moreover, we included a sub-task that prompts language models to classify also the fallacy type before repairing it in configurations where the fallacy label is not provided. By comparing the models' performance when the fallacy la-

bel is given as input versus when it is omitted, we can evaluate both their ability in fallacy classification and in the repairing task, in order to provide a more comprehensive assessment of the models' capabilities in handling fallacious arguments under different scenarios. Our research questions are the following:

**RQ1**: Can LLMs accurately classify the labels of fallacious arguments in *configurations* where the fallacy label is not provided?

**RQ2**: How effectively can LLMs *repair* fallacious arguments in political debates into clearer, fairer statements devoid of manipulative rhetoric, compared to human annotations?

**RQ3**: Beside enhancing human comprehension, do the *repaired* arguments also provide deeper insights on the original input message?

To address these questions, we considered various strategies, detailed in Section 4.1, and provide different input to the model during the prediction and/or repairing process.

### 4.1 Prompt Construction Methodology

The LLMs generation process is guided by a carefully tailored instruction prompt, designed to direct LLMs in generating outputs that can be compared to the gold standard expert arguments previously annotated. We tested various prompt configurations by including or excluding two key elements: *(1)* the fallacy label (`L`) and *(2)* the contextual information (`C`) surrounding the fallacy to be repaired. We conducted our experiments under the Zero-Shot, Few-Shot and Fine-Tuning settings for all the configurations: `LC` (fallacy label and context), `CO` (context only), `LO` (fallacy label only), and `NO` (no fallacy label and context). For the `LO` and `NO` configurations, we utilized a subset of the FallacyFix dataset. This subset specifically included 541 examples where the fallacies were repaired through paraphrasing or partial text modification. We excluded cases where the repair method involved complete removal of the original fallacious text. This selection criterion was implemented because providing a fallacious snippet for classification in the prompt becomes impractical when the repaired version has entirely eliminated the original text.

The prompt configuration illustrated in Figure 4 employs a modular approach, utilizing color-coded components to delineate distinct functional roles.

The **orange** section provides context when present, while the core fallacious argument in **pine green** forms the base of analysis. Fallacy label handling is managed through mutually exclusive components: the **green** section is used when the label is provided, while the **cyan** section requests a label for the fallacy classification subtask. The prompt's complexity scales across different configurations: *Zero-Shot* uses only base components, *Few-Shot* (**magenta**) adds five demonstrative examples, and

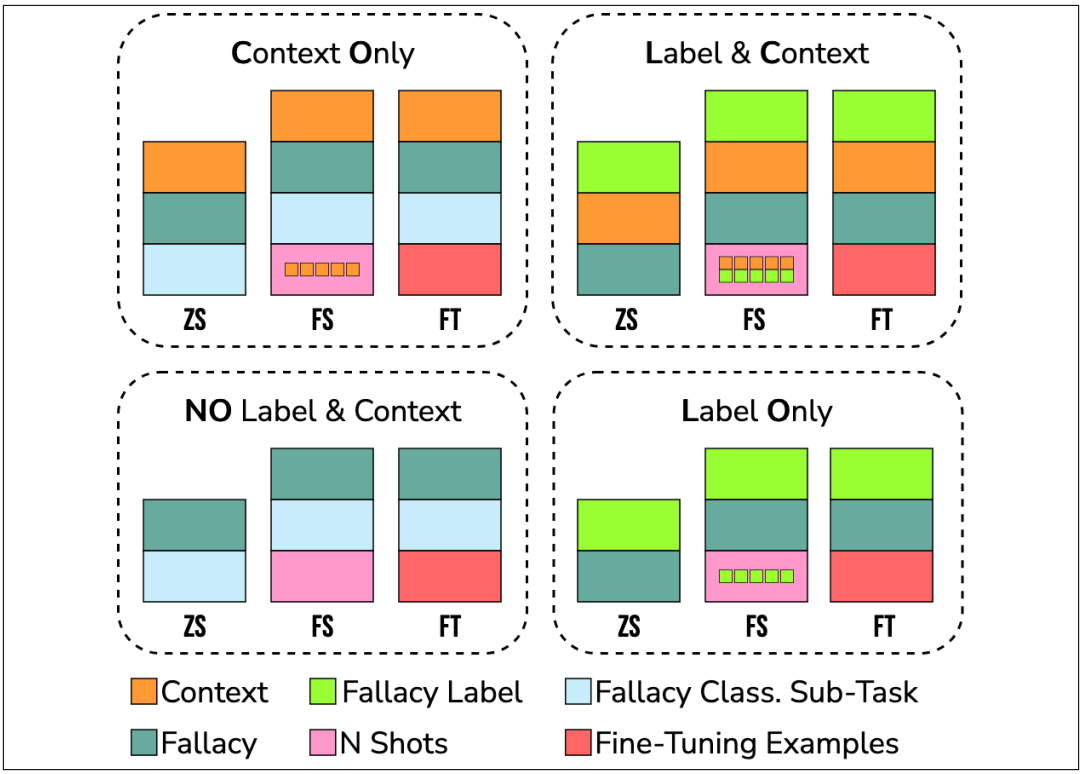


Figure 4: Prompt modularity based on the specific configuration and setting.

*Fine-Tuning* (**red**) incorporates additional training instructions.

### 4.2 Method

To address the aforementioned tasks, we employed several widely used LLMs and a baseline model for comparison. The evaluated models include open-source options: BART [Lewis et al., 2019], Google Gemma [Team et al., 2024], Mistral 7B [Jiang et al., 2023], Mixtral 8x7B [Jiang et al., 2024], and LLaMa 3 [AI@Meta, 2024], all accessed via Hugging Face. We also assessed proprietary models: OpenAI GPT (*gpt-3.5-turbo-0125* and *gpt-4-0125*) and Claude [Anthropic, 2023] (*claude-3-opus-20240229*), accessed through their respective APIs[7].

[7]Exploration and testing of LLMs were concluded on 30/04/2024.

### 4.3 Metrics

**Metrics for the automatic evaluation.** We employ standard text completion metrics to evaluate the similarities and differences between the generated text and the gold-standard repaired text. Since fallacious arguments can be repaired through various methods such as paraphrasing, omission, generalization, or weakening, we assess the extent to which the generated text aligns with the human annotators' repaired arguments. Different word combinations can often convey the same meaning. Therefore, we focus on both lexical overlap and semantic similarity. This approach ensures that the generated text maintains the original argument's core intent and clarity.

**BERTScore** [Zhang et al., 2020] evaluates the similarity between generated and reference texts using contextual embeddings from the BERT model to capture semantic meaning.[8]

**IOU-F1** measures the overlap between the predicted and true labels, combining the Intersection over Union (IoU) metric with the F1 score to assess both precision and recall.

**Macro AVG F1 Score for Fallacy Classification** provides a balanced assessment of model performance across all fallacy labels, ensuring that no single class dominates the evaluation, especially in the presence of class imbalances.

We considered these metrics as the most relevant ones for our study due to their robust methodologies, alignment with the task objectives, and the meaningful insights they provide into the results.

**Metrics for human evaluation.** Based on prior research [Sourati et al., 2023, Clinciu et al., 2021, Wang et al., 2023, Wang and Dong, 2020], we evaluated the unveiled fallacious arguments for Relevance, Suitability, and Cogency to assess if they are clear and comprehensible, and to identify the optimal approach for paraphrasing the fallacies according to the LLMs' results using a Likert scale from 1 to 5 for every metric.

**Relevance**: The repaired fallacious argument aligns with the topic and category of the original fallacy.

**Suitability**: The style of the repaired fallacy is appropriate, maintaining politeness, neutrality, and avoiding explicit references while expressing the same meaning.

**Cogency**: The repaired fallacy demonstrates logical correctness and coherence, w.r.t. the prompt.

To the best of our knowledge, no other work has employed these evaluation techniques in the context of addressing fallacious arguments in political debates. By

[8] `https://huggingface.co/docs/evaluate/`

adapting these methods to the unique challenges of this domain, we aim to establish a benchmark for future research and contribute to the advancement of evaluation techniques tailored to repair fallacies in political debates.

### 4.4 Experimental Settings

As described in Section 4.1, we conducted experiments using different configurations. We employed the Low-Rank Adaptation (LoRA) [Hu et al., 2021] technique for our experiments, which allowed us to efficiently load and fine-tune large language models without exceeding hardware limitations. We maintained consistent parameters across all LLMs to ensure comparable results. In both *Zero-Shot* and *Few-Shot* settings, we set the `temperature` to `0.5` to balance creativity and logical coherence, and the `max_tokens` to `512` to ensure concise responses. To balance computational efficiency and LLM capability exploration, we fixed the number of responses to `1`. The `stop` parameter was left unset (`None`), allowing the models to determine the natural endpoint of each response autonomously. In the *Few-Shot* setting, we randomly sampled one example for each of the five fallacy subcategories involved in the experiment. The number of demonstration examples was chosen based on pilot experiments, which indicated that it balances both efficiency and quality. This aligns with prior research in the hate speech domain showing that providing 5–10 examples is generally sufficient for LLMs [Ocampo et al., 2023, Hartvigsen et al., 2022]. The connection between hate speech and fallacious arguments stems from both phenomena involving the use of biased or deceptive language to unfairly influence opinions or target certain groups [Meade, 2021]. Throughout the *Fine-Tuning* process, we employed additional parameters to facilitate our experiments. The `max_seq_length` parameter was set to `1024` to optimize the length of both input and output texts. We set `3` epochs of training, with a `learning_rate` of `2e-4` and a `weight_decay` of `0.01`. Utilizing the `paged_adamw_32bit` optimizer, we configured the train and evaluation `batch` sizes as `4` and `2`, respectively.

## 5 Evaluation

To address **RQ1**, we evaluated LLMs' ability to categorize fallacies under *Zero-Shot*, *Few-Shot*, and *Fine-Tuned* settings. Our dual-task framework highlights LLMs' versatility in jointly performing fallacy classification and text generation, setting our work apart from prior classification-focused studies Goffredo et al. [2022], Habernal et al. [2018a], Jin et al. [2022], Vijayaraghavan and Vosoughi [2022]. Table 2 reports macro F1 scores for Context Only (`CO`) and No Fallacy Label & Context (`NO`) configurations. GPT-4 performs best in ZS and FS, reflecting strong reasoning abilities.

LLaMA 3 8B leads in FT showing effective adaptation.

| | Context Only (CO) | | | | | | No Fall. label & Context (NO) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ZS | | FS | | FT | | ZS | | FS | | FT | |
| Model | F1 | #PL | F1 | #PL | F1 | #PL | F1 | #PL | F1 | #PL | F1 | #PL |
| *BART* | - | - | - | - | *34,92%* | *5* | - | - | - | - | *42,20%* | *5* |
| Claude 3 | 37,94% | 7 | 37,22% | 7 | - | - | 18,55% | 13 | 23,41% | 9 | - | - |
| Gemma 1.1 2B | 6,23% | 8 | 2,35% | 172 | 3,70% | 8 | 5,22% | 7 | 1,44% | 4 | 4,88% | 5 |
| Gemma 1.1 7B | 2,36% | 71 | 0,91% | 5 | 15,58% | 14 | 1,54% | 70 | 0,84% | 99 | 6,69% | 13 |
| GPT-3.5 | 12,74% | 19 | 26,68% | 9 | 22,95% | 10 | 7,44% | 23 | 30,52% | 6 | 13,33% | 10 |
| GPT-4 | 43,48% | 7 | **59,15%** | 5 | - | - | 26,32% | 9 | **37,69%** | 6 | - | - |
| LLaMa 3 8B | 6,68% | 34 | 2,80% | 7 | 52,76% | 5 | 6,64% | 23 | 6,82% | 25 | 31,06% | 2 |
| Mistral 7B | 0,00% | 368 | 0,44% | 264 | - | - | 0,54% | 156 | 0,59% | 182 | - | - |
| Mixtral 8x7B | 0,79% | 193 | 1,05% | 159 | 12,96% | 12 | 0,92% | 158 | 1,34% | 112 | 4,48% | 14 |

Table 2: Macro F1 Score and #Predicted Labels (#PL) results for all the models in every setting and configuration.

To answer to **RQ2**, we compared the generated arguments from the LLMs with the gold standard described in Section 3.1. We utilized BERTScore and IOU-F1 as main metrics for semantic-based automatic evaluation and token-based, respectively, to assess performance, using the same approach described in Section 3.2. Our first analysis revealed that human-repaired fallacies closely match the original arguments, with a BERTScore of $0.91 \pm 0.06$. Different human repairs showed even higher similarity ($0.94 \pm 0.06$). Notably, human repairs outperformed those by LLMs when compared to the original fallacies, suggesting that AI systems still lag behind humans in this complex task. Furthermore, as summarized in Table 3, LLaMa 3 consistently achieves the best scores across all configurations, with average scores of 0,94 and 0,97 for BERTScore and IOU-F1, respectively. Moreover, in the Fine-Tuning (FT) setting, the metrics show significantly higher results for configurations that omit the fallacy context (`FO`, `NO`) compared to the Zero-Shot (ZS) and Few-Shot (FS) settings. In both ZS and FS settings, configurations without fallacy context (`LO`, `NO`) result in lower metric values. In contrast, configurations with context (`CO`, `LC`) exhibit higher metric values, indicating that models perform better with the presence of context, even without pre-training. In addition, we also reported the results of BART in the Fine-Tuning settings, where it achieves the best result in both of the configuration (`CO` and `NO`)[9] with the aim to establish a baseline model to compare with.

After identifying the best models, we conducted a human annotation study using the metrics detailed in Section 4.3. This study addressed **RQ3**, where 17 annotators voluntarily evaluated 15 repaired arguments generated by the top models in each setting and configurations. The annotation process was carried out in a controlled

[9]The method used with BART for fallacy label classification and text generation differs significantly from other LLMs due to its unique prompting requirements.

environment[10].

To assess inter-rater reliability, we used Krippendorff's $\alpha$ [Krippendorff, 2011], calculated based on all the repaired arguments annotated by the 17 annotators. The values ranged from 0.16 to 0.19 across all criteria, indicating low reliability.

The percentage agreements ranged from 49% to 60%, offering a somewhat more positive outlook. Suitableness had the highest agreement at 60%, while Cogency had the lowest at 49%. In terms of ratings, LLM-generated annotations were deemed relevant (mean score of $4.03 \pm 0.68$) and suitable ($4.17 \pm 0.68$) but were rated lower in Cogency ($3.76 \pm 0.69$) on a 5-point scale. Specifically, Relevance showed a Krippendorff's $\alpha$ of 0.16 with 55% agreement, Suitableness had an $\alpha$ of 0.19 with 60% agreement, and Cogency also had an $\alpha$ of 0.19 but with 49% agreement.

The annotator demographics show an unbalanced distribution. There is a male majority (64.7%), and most participants (76.5%) are aged 21-30.

[10]On-site collaboration between researchers and annotators enabled real-time communication, improving annotation consistency and quality through prompt feedback and quick problem-solving.

| | *Zero-Shot* | | | | | | | | *Few-Shot* | | | | | | | | *Fine-Tuning* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BERTScore | | | | IOU F1 | | | | BERTScore | | | | IOU F1 | | | | BERTScore | | | | IOU F1 | | | |
| **Model** | CO | LC | NO | LO | CO | LC | NO | LO | CO | LC | NO | LO | CO | LC | NO | LO | CO | LC | NO | LO | CO | LC | NO | LO |
| *BART* | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | *0,98* | *0,98* | - | - | *0,99* | *0,98* | - | - |
| Claude 3 | 0,68 | 0,69 | 0,52 | 0,54 | 0,89 | 0,89 | 0,75 | 0,76 | 0,71 | 0,78 | 0,64 | 0,65 | 0,89 | 0,92 | 0,80 | 0,80 | - | - | - | - | - | - | - | - |
| Gemma 1.1 2B | 0,62 | 0,53 | 0,64 | 0,49 | 0,86 | 0,85 | 0,80 | 0,74 | 0,38 | 0,49 | 0,39 | 0,49 | 0,61 | 0,83 | 0,62 | 0,75 | 0,49 | 0,49 | 0,55 | 0,48 | 0,76 | 0,77 | 0,80 | 0,72 |
| Gemma 1.1 7B | 0,55 | 0,54 | 0,52 | 0,50 | 0,85 | 0,85 | 0,75 | 0,74 | 0,51 | 0,61 | 0,49 | 0,72 | 0,83 | 0,82 | 0,75 | 0,83 | 0,51 | 0,51 | 0,51 | 0,49 | 0,78 | 0,77 | 0,74 | 0,73 |
| GPT 3.5 turbo | 0,68 | 0,70 | 0,62 | 0,59 | 0,88 | 0,89 | 0,79 | 0,78 | 0,65 | - | 0,69 | 0,62 | 0,84 | - | 0,81 | 0,81 | 0,68 | 0,69 | 0,66 | 0,63 | 0,88 | 0,89 | 0,82 | 0,82 |
| GPT 4 | 0,69 | 0,71 | 0,61 | 0,60 | 0,89 | 0,90 | 0,79 | 0,74 | 0,70 | - | 0,66 | - | 0,89 | - | 0,80 | - | - | - | - | - | - | - | - | - |
| LLaMa 3 8B | 0,66 | 0,67 | 0,56 | 0,57 | 0,88 | 0,88 | 0,76 | 0,76 | 0,70 | 0,55 | 0,60 | 0,52 | 0,88 | 0,79 | 0,78 | 0,68 | 0,93 | 0,88 | 0,96 | **0,97** | 0,97 | 0,97 | 0,96 | **0,98** |
| Mistral 7B | 0,58 | 0,58 | 0,53 | 0,53 | 0,78 | 0,81 | 0,69 | 0,69 | 0,58 | 0,61 | 0,58 | 0,54 | 0,78 | 0,85 | 0,78 | 0,71 | - | - | - | - | - | - | - | - |
| Mixtral 8x7B | 0,60 | 0,62 | 0,57 | 0,53 | 0,81 | 0,82 | 0,73 | 0,71 | 0,60 | 0,43 | 0,59 | 0,43 | 0,83 | 0,74 | 0,75 | 0,72 | 0,76 | 0,79 | 0,62 | 0,58 | 0,92 | 0,92 | 0,81 | 0,78 |

Table 3: Full table with the results of BERTScore and IOU F1 in every setting and configuration.

Educational backgrounds are diverse. The majority are PhD students (41.2%) and Master's degree holders (35.3%). Smaller proportions include researchers (17.6-%) and Bachelor's degree holders (5.9%).

After addressing **RQ3**, our error analysis focused on inter-rater reliability in evaluating LLMs outputs. We observed high percentage agreements among annotators, suggesting that LLMs generally produce content that is fitting and relevant. However, this analysis also revealed a high degree of subjectivity in evaluations, with annotators often reaching similar judgments through different reasoning. This finding underscores the need for developing specific, standardized metrics tailored to human evaluation tasks. Such metrics would not only ensure greater consistency and reliability in assessments but also guide annotators towards more objective evaluations. Furthermore, while the high agreement rates are encouraging, they indicate that LLMs outputs require improvement in coherence to enhance their overall quality due to a discrete number of "% Not Matched" answers.

## 6 Conclusion

Detecting and addressing fallacies is a highly complex task [Van Bouwel, 2003], particularly in the nuanced context of political debates. Disagreement over whether a fallacy should be corrected, along with the often blurred boundaries of what truly constitutes a fallacy, further complicates this challenge. This research presents four significant contributions to the field of fallacious argument repair. Firstly, we introduce FallacyFix, a novel dataset of repaired arguments from fallacies. Secondly, we demonstrate the effectiveness of simultaneously predicting fallacy labels and repairing arguments using this dataset. Thirdly, our analysis of generated fallacy-repairing arguments shows promise in Few-Shot settings with contextual prompts, but reveals limitations in Zero-Shot and Fine-Tuning approaches, challenging the identification of an optimal model for accurate fallacy repair in political debates. Finally, an extensive user study validates the relevance, suitability, and cogency of the repaired arguments, underscoring the practical utility of our approach. The module to repare fallacious argument described in this paper has been integrated into the web application DispuTOOL [Goffredo et al., 2025], which enables users to visualize argument components and their boundaries, explore the relations between these components, and examine fallacious arguments alongside their repaired versions. Future research will focus on giving a more specific definition of each fallacy, integrating domain-specific knowledge to address complex fallacy categories, further analyzing language models' behavior in countering fallacies, and exploring real-time fallacy repair methodologies. These efforts aim to enhance our ability to address

fallacies dynamically in various argumentation contexts, potentially improving the quality of public discourse and decision-making.

## References

AI@Meta. The llama 3 foundational language model. `https://ai.meta.com/blog/meta-llama-3/`, 2024. Accessed: 31 May 2024.

Anthropic. Introducing the claude 3 family of ai models. `https://www.anthropic.com/news/claude-3-family`, 2023. Accessed: 31 May 2024.

Leticia Bode and Emily Vraga. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65, 08 2015. doi: 10.1111/jcom.12166.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation*, 29(4):431–456, 2015.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. Exploring the potential of large language models in computational argumentation, 2024.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. A study of automatic metrics for the evaluation of natural language explanations. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online, April 2021. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. 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 5636–5646, 2019.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Frans H. Van Eemeren. Fallacies. *Critical concepts in argumentation theory*, page 135–164, 2001.

Frans H. Van Eemeren and Rob Grootendorst. Fallacies in pragma-dialectical perspective. *Argumentation*, 1(3):283–301, 1987. doi: 10.1007/bf00136779.

Frans H. van Eemeren and Rob Grootendorst. *A Systematic Theory of Argumentation: The pragma-dialectical approach*. Cambridge University Press, 2003.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization, 7 2022.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. Argument-based detection and classification of fallacies in political debates. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore, December 2023. Association for Computational Linguistics.

Pierpaolo Goffredo, Deborah Dore, Elena Cabrio, and Serena Villata. DISPUTool 3.0: Fallacy detection and repairing in argumentative political debates. In Pushkar Mishra, Smaranda Muresan, and Tao Yu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 472–480, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.45. URL https://aclanthology.org/2025.acl-demo.45/.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018a.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics.

Charles L. Hamblin. Fallacies. *Tijdschrift Voor Filosofie*, 33(1):183–188, 1970.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. Mafalda: A benchmark and comprehensive study of fallacy detection and classification, 2024.

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

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. Mistral 7b, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*, 2022.

Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

Stephan Lewandowsky, Ullrich Ecker, Colleen Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13:106–131, 12 2012. doi: 10.1177/1529100612451018.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding, 2024a.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. When llms meet cunning questions: A fallacy understanding benchmark for large language models, 2024b.

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

Lynn Meade. Fallacies–warning! deceptive, hateful speech coming your way. *Advanced Public Speaking*, 2021.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2758–2772, Toronto, Canada, July 2023. Association for Computational Linguistics.

Steve Oswald and Thierry Herman. *Give the Standard Treatment of Fallacies a Chance! Cognitive and Rhetorical Insights into Fallacy Processing*, pages 41–62. Springer International Publishing, Cham, 2020.

F Paglieri. The scaremongering fallacy of fallacy theory: How to improve reasoning without fear of error. *Natural Arguments: A Tribute to John Woods, London: College Publications*, pages 79–101, 2019.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. How susceptible are LLMs to logical fallacies? In Nicoletta Calzolari, Min-Yen

Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8276–8286, Torino, Italia, May 2024. ELRA and ICCL.

Alan Ramponi, Agnese Daffara, and Sara Tonelli. Fine-grained fallacy detection with human label variation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 762–784, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.34. URL https://aclanthology.org/2025.naacl-long.34/.

Ramon Ruiz-Dolz and John Lawrence. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In Milad Alshomary, Chung-Chi Chen, Smaranda Muresan, Joonsuk Park, and Julia Romberg, editors, *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore, December 2023. Association for Computational Linguistics.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments, 2023.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43, 04 2016.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Jeroen Van Bouwel. When unveiling the epistemic fallacy ends with committing the ontological fallacy. on the contribution of critical realism to the social scientific explanatory practice. *Philosophica*, 71(1), 2003.

Prashanth Vijayaraghavan and Soroush Vosoughi. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States, July 2022. Association for Computational Linguistics.

Jacky Visser, John Lawrence, and Chris Reed. Reason-checking fake news. *Commun. ACM*, 63(11):38–40, 2020.

Douglas Walton. *Argumentation Schemes*. Cambridge University Press, 2008.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. Evaluating gpt-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6255–6263, 2023.

Jiapeng Wang and Yihong Dong. Measurement of text similarity: A survey. *Information*, 11(9), 2020. ISSN 2078-2489.

John Woods. *Errors of reasoning: Naturalizing the logic of inference*. College Publ., 2013.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

Received \jreceived