2/9/2026 1:04:03 PM

# Compare Results

| Old File: | versus | New File: |
|---|---|---|
| **2024.emnlp-main.730.pdf** | | **2024_emnlp-main_730.pdf** |
| **17 pages (15.32 MB)** | | **20 pages (202 KB)** |
| | | 2/8/2026 5:29:46 AM |

**Total Changes**

37

**Content**

2 Replacements

19 Insertions

16 Deletions

**Styling and Annotations**

0 Styling

0 Annotations

Go to First Change (page 1)

# Boosting Logical Fallacy Reasoning in LLMs via Logical Structure Tree

Yuanyuan Lei and Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University, College Station, TX
{yuanyuan, huangrh}@tamu.edu

February 8, 2026

**Abstract**

Logical fallacy uses invalid or faulty reasoning in the construction of a statement. Despite the prevalence and harmfulness of logical fallacies, detecting and classifying logical fallacies still remains a challenging task. We observe that logical fallacies often use connective words to indicate an intended logical relation between two arguments, while the argument semantics does not actually support the logical relation. Inspired by this observation, we propose to build a logical structure tree to explicitly represent and track the hierarchical logic flow among relation connectives and their arguments in a statement. Specifically, this logical structure tree is constructed in an unsupervised manner guided by the constituency tree and a taxonomy of connectives for ten common logical relations, with relation connectives as non-terminal nodes and textual arguments as terminal nodes, and the latter are mostly elementary discourse units. We further develop two strategies to incorporate the logical structure tree into LLMs for fallacy reasoning. Firstly, we transform the tree into natural language descriptions and feed the textualized tree into LLMs as a part of the hard text prompt. Secondly, we derive a relation-aware tree embedding and insert the tree embedding into LLMs as a soft prompt. Experiments on benchmark datasets demonstrate that our approach based on logical structure tree significantly improves precision and recall for both fallacy detection and fallacy classification.

## 1 Introduction

Logical fallacy refers to the use of invalid or flawed reasoning in an argumentation Risen2007,Walton2010,Cotton2018. Logical fallacy can occur as unintentional mistakes or deliberate persuasions in a variety of human communications, such as news media DaSanMartino2019, educational essay Jin2022, political debates Goffredo2023,Mancini2024, or online discussions Sahai2021. Logical fallacies can lead to harmful consequences for society, such as spreading misinformation Musi2022,Lundy2023, raising public health risks Lin2020, manipulating public opinions Barclay2018,Lei2022,Lei2024a, introducing societal bias and polarization AbdEldayem2023. Despite their prevalence and harmfulness, understanding logical fallacies still remains a challenging task, which requires both semantics understanding and logical reasoning Li2022,Sanyal2023. In this paper, we focus on fallacy detection and classification, and aim to develop an approach that generalizes across different domains and genres.

The key observation is that logical fallacies heavily rely on connective phrases to indicate an intended logical relation between two textual arguments, while the semantics of the arguments do not actually support the claimed logical relation. Figure **??** shows two examples where the connective phrases were bolded. The first example uses the connective words *therefore* and *cause* to suggest a causal relation between vaccinations and increasing flu cases, however, the temporal relation between the two events as stated in the first half of the statement does not necessarily entail a causal relation between them, and indeed, their semantics do not actually support the suggested causal relation. Recognizing this discrepancy undermines the credibility of the whole statement. Similarly in the second example, the connective word *likewise* is commonly used to indicate an analogy relation, however, the second argument is clearly a specific case of the general condition stated in the first argument and therefore there is no analogy relation between them, and recognizing this mismatch between the suggested logical relation and the real relation enables us to detect this fallacy.

Therefore, we propose to construct a logical structure tree that organizes all connective phrases in a statement and their textual arguments into a hierarchical structure. We expect the logical structure tree to effectively capture the juxtaposition of connective phrase suggested logical relations and the real logical relations between textual arguments, and therefore guide LLMs in fallacy detection and classification. Specifically, a logical structure tree consists of relation connectives as non-terminal nodes and textual arguments as terminal nodes, and the latter mostly corresponds to elementary discourse units (EDU) considered in discourse parsing. Figure **??** shows the logical structure trees constructed for the two example texts.

As the logical relation indicated by a connective phrase may not be supported by semantics of its arguments in the context, we identify the purposefully indicated logical relations in a context-free unsupervised manner by matching a connective phrase with a taxonomy of connectives compiled for ten common logical relations (conjunction, alternative, restatement, instantiation, contrast, concession, analogy, temporal, condition, causal). To construct a logical structure tree, we first construct a constituency tree for a statement and then search in the constituency tree for connective phrases in the top-down left to right order, and the first found connective phrase will be the root node of the logical structure tree. Next, we identify the text spans of its two arguments using rules and recursively build the left and right sub-trees by applying the same procedure to constituency tree segments corresponding to the two arguments.

The logical structure tree is integrated into LLMs for fallacy reasoning using two strategies. The first considers textualized tree, where we convert the tree into natural language descriptions, making the tree readable by LLMs. Particularly, we describe the relations and arguments in a bottom-up manner, providing the LLMs with insight into logical relations from a local to global perspective. We then concatenate the textualized tree with the instruction prompt, and input them into LLMs as a hard prompt. The second considers tree-based soft prompt, where we derive a relation-aware tree embedding. Specifically, we design relation-specific encoders to process each type of relation and incrementally derive the tree embedding from bottom up to the root node. We then insert the tree embedding into LLMs as a soft prompt for further tuning.

Experiments on benchmark datasets across various domains and genres validate that our approach based on logical structure tree effectively improve precision and recall for both fallacy detection and fallacy classification tasks. Our main contributions are summarized as follows:

- We propose to construct a logical structure tree to capture the juxtaposition of connective phrase suggested logical relations and the real logical relations between textual arguments, and use it to serve as additional guidance for fallacy detection and classification.
- We effectively improve the F1 score for fallacy detection by up to 3.45% and fallacy classification by up to 6.75% across various datasets.

## 2 Related Work

**Logical Fallacy** is erroneous patterns of reasoning Walton1987,Fantino2003. Initial work explored the taxonomy of fallacies Tindale2007,Greenwell2006,Walton2008. Recent works have focused on the automatic detection and classification of fallacies. Habernal2017 developed a software that deals with fallacies in question-answering. Sheng2021 investigated ad hominem fallacy in dialogue responses. Habernal2018 explored the ad hominem fallacy from web argumentations. Stab2017 recognized insufficient arguments in argumentation essays. Goffredo2022 categorized fallacies in political debates. Nakpih2020 focused on fallacies in legal argumentations. Musi2022 researched fallacies about pandemics on social medias. Alhindi2022 proposed a multi-task prompting approach to learn the fallacies from multiple datasets jointly. Jin2022 proposed a structure-aware method to classify fallacies. Different from Jin2022 that masked out content words to form a sequence-based pattern, our paper proposes a tree-based hierarchical logical structure to unify both relation connectives and content arguments together.

**Logical Reasoning** abilities of large language models are gaining increasing research attention Xu2023,Chen2021,Creswell2022,Pi20 Olausson2023 combined large language models with first-order logic. Pan2023,Zhang2023 empowered large language models with symbolic solvers. Pi2022 presented an adversarial pre-training framework to improve logical reasoning. Zhao2023 incorporated multi-step explicit planning into the inference procedure. Jiao2022 proposed a contrastive learning approach to improve logical question-answering. Different from these previous work, we particularly focus on logical fallacy reasoning, aiming to detect and classify fallacies.

**Misinformation** refers to the unverified or false information Guess2020,Armitage2021,Aimeur2023,Lei2024b. Misinformation detection was studied for years, such as fake news Rashkin2017,Lei2023b,Oshikawa2020, rumor

**[False Cause]** The region continues to report flu incidents after many people took the vaccination, therefore, the vaccinations cause increasing flu cases.

```
          [therefore (causal)]
         /                    \
  [after (temporal)]      the vaccinations cause increasing flu cases
   /              \
the region    many people
continues     took the
to report     vaccination
flu incidents
```

**[Hasty Generalization]** People will never get ill as long as they take this pill every day, likewise, my sister takes it regularly and is always healthy.

```
            [likewise (analogy)]
           /                    \
  [as long as (condition)]    [and (conjunction)]
      /            \           /                \
People will   they take   my sister      is always
never get     this pill   takes it      healthy
ill           every day   regularly
```

Figure 1: Examples of logical fallacy sentences and their logical structure trees. The logical structure tree features logical relation connectives as non-terminal nodes, and textual arguments as terminal nodes.

Ma2018,Li2019, satire Yang2017, political bias Lei2022,Feng2023,Devatmane2023,Lei2024, propaganda DaSanMartino2019,DaSanMartino2020,Lei2023a. Logical fallacies are often employed within misinformation to present invalid claim as credible, facilitating the spread of misinformation Beisecker2024,Pauli2022,Bonial2022. Developing automatic models to detect logical fallacies can also benefit the identification and mitigation of misinformation.

# 3 Logical Structure Tree

The logical structure tree consists of relation connectives as non-terminal nodes, and textual arguments as terminal nodes. The relation connectives serve as parent nodes, and the two corresponding arguments are linked as left and right children nodes. Figure **??** illustrates examples of the logical structure tree. The logical structure tree is constructed in an unsupervised manner, guided by the constituency tree and a taxonomy of connectives compiled for ten common logical relations.

## 3.1 Relation Connectives

The logical fallacies usually rely on relation connectives to indicate a logical relation. Inspired by the discourse relations proposed by Prasad2008, we define a taxonomy of ten logical relations which are commonly seen: conjunction, alternative, restatement, instantiation, contrast, concession, analogy, temporal, condition, and causal relations. Moreover, we build a set of connective words and phrases that correspond to each type of logical relation, as shown in Table **??**. This set of connectives includes the explicit discourse connectives from the PDTB discourse relation dataset Prasad2008, and is further expanded by manually adding relevant connectives from the development set of the logic fallacy dataset Jin2022.

We further conduct a statistical analysis on the distribution of ten logical relations and compare distributions between fallacy and no fallacy classes as well as across different fallacy classes, with the detailed results shown in Appendix A. The statistical analysis shows that both the fallacy and no fallacy classes contain many connective phrases and their distributions of the ten logical relations are also very similar. But as expected, different fallacy types tend to employ varying logical patterns, for example, False Dilemma uses more alternative relation, while Deductive Fallacy uses more analogy relation.

Table 1: The ten types of logical relations and their relation connectives.

| Logical Relations | Relation Connectives |
|---|---|
| conjunction | and, as well as, as well, also, separately |
| alternative | or, either, instead, alternatively, else, nor, neither |
| restatement | specifically, particularly, in particular, besides, additionally, in addition, moreover, furthermore, plus, not only, indeed, in other words, in fact, in short, in the end, overall, in summary, in details |
| instantiation | for example, for instance, such as, including, as an example, as an instance, for one thing |
| contrast | but, however, yet, while, unlike, rather, rather than, in comparison, by comparison, on the other hand, on the contrary, contrary to, in contrast, by contrast, whereas, conversely, not, no, none, nothing, n't |
| concession | although, though, despite, despite of, in spite of, regardless, regardless of, nevertheless, nonetheless, even if, even though, even as, even when, even after, even so, no matter |
| analogy | likewise, similarly, as if, as though, just as, just like, namely |
| temporal | during, before, after, when, as soon as, then, next, until, till, meanwhile, in turn, meantime, afterwards, simultaneously, at the same time, beforehand, previously, earlier, later, thereafter, finally, ultimately |
| condition | if, as long as, unless, otherwise, except, whenever, whichever, once, only if, only when, depend on |
| causal | because, cause, as a result, result in, due to, therefore, hence, thus, thereby, since, now that, consequently, in consequence, in order to, so as to, so that, why, for, accordingly, given, turn out |

### 3.2 Tree Construction Algorithm

To construct a logical structure tree $T_{logic}$, we first construct a constituency tree $T_{con}$ for a statement. We use the stanza library[1] to get the constituency tree Qi2020. At the beginning, $T_{logic}$ is initialized as an empty tree. Then we traverse the constituency tree $T_{con}$ from top to bottom and from left to right, and match relation connectives within each subtree of $T_{con}$. If there is a subtree $S_{con}(w)$ whose text equals to a relation connective $w$, we use the algorithm in section **??** to extract the two textual arguments $\alpha, \beta$ associated with $w$. Then a new logical subtree $S_{logic}(w)$ is created, with the matched relation connective $w$ as a parent node, and the two arguments $\alpha, \beta$ as its left and right children. This new logical subtree $S_{logic}(w)$ is added into the logical structure tree $T_{logic}$. If the textual arguments $\alpha, \beta$ still contain other relation connectives, then we recursively match relation connectives in the arguments and replace the original argument node in the $T_{logic}$ with the newly created logical subtree. The termination condition is that all the relation connectives in the given text have been matched.

### 3.3 Textual Arguments Extraction

The textual arguments are the two content components linked by a relation connective. Given a matched relation connective $w$, its corresponding subtree in the $T_{con}$ is $S_{con}(w)$. To extract the arguments of $w$, we find the parent tree of $S_{con}(w)$ in the $T_{con}$, denoted as $P(S_{con}(w))$. The text enclosed by $P(S_{con}(w))$ is the concatenation of all its leaf node texts. If the text enclosed by parent tree $P(S_{con}(w))$ contains content before and after the relation connective $w$, i.e., has the form of $\alpha + w + \beta$, then the left argument of $w$ is $\alpha$ and the right argument is $\beta$. If the text enclosed by parent tree $P(S_{con}(w))$ only contains content after the relation connective $w$, i.e., has the form of $w + \beta$, then the right argument of $w$ is $\beta$, and the left argument $\alpha$ is the text enclosed by grandparent tree $P(P(S_{con}(w)))$ subtracted by the text enclosed by $P(S_{con}(w))$.

## 4 Logical Fallacy Reasoning

We further design a framework to incorporate the logical structure tree into LLMs for fallacy detection and classification. This framework consists of two main components. The first is textualized tree, where we convert the logical structure tree into natural language descriptions, and feed it into LLMs as a hard text prompt. The second is tree-based soft prompt, where we derive a relation-aware tree embedding, and insert it into LLMs as a soft prompt for additional tuning. The hard and soft prompts are complementary: the hard prompt enriches the instruction with logical structure information, while the soft prompt facilitates direct tuning on tree embeddings. Figure **??** shows an illustration.

[1] https://stanfordnlp.github.io/stanza/constituency.html

### 4.1 Textualized Tree

The textualized tree aims to transform the logical structure tree into the textual form, which can be interpretable by LLMs. As shown by the upper path of Figure **??**, the textualized tree is represented as a table which consists of three columns: left argument, relation connective, right argument. Each row in the table represents a triplet (left argument, relation connective, right argument) corresponding to each logical relation in the tree. In particular, we organize the triplets into the table in a bottom-up order, to provide the LLMs with insight into logical relations from a micro to macro perspective. The textualized tree is then input into the LLMs as a part of the hard text prompt:

$$h_t = \text{TextEmbedder}(\text{textualize}(T_{logic})) \tag{1}$$

where $\text{textualize}(\cdot)$ denotes the textualization operation, TextEmbedder refers to the text embedding layer of LLMs, $h_t$ is the mapped embedding of the textualized tree.

### 4.2 Tree-based Soft Prompt

The tree-based soft prompt is a tree embedding which is projected into LLMs as a soft prompt for further tuning. As shown by the lower path of Figure **??**, this process includes a tree encoder to derive the tree embedding, as well as a projection layer to transform the tree embedding into the representation space of LLMs.

During the tree encoder stage, we aim to derive a relation-aware tree embedding. To integrate relation information into tree embedding, we design relation-specific encoders to process each type of logical relation. For a simple tree whose children nodes are leaf nodes without hierarchical layers, its embedding is computed as:

$$e_r = W^r(e_\alpha \oplus e_c \oplus e_\beta) + b^r \tag{2}$$

where $e_r$ is the embedding of this simple tree, $e_\alpha$, $e_c$, $e_\beta$ are the embeddings of left argument, relation connective, and right argument, which are initialized as the average of word embeddings derived from RoBERTa language model Liu2019; $\oplus$ denotes feature concatenation; $W^r$, $b^r$ are the trainable parameters of the encoder that corresponds to the relation type $r$, where $W^r \in \mathbb{R}^{d\times 3d}$, $b^r \in \mathbb{R}^d$, and $d = 768$ is the dimension of embedding space in RoBERTa. The relation type $r$ is one of the ten logical relations associated with the relation connective.

For the tree with hierarchical structure, we derive the tree embedding incrementally, starting from the bottom simple tree and up towards the root node:

$$e_r = W^r(e_l \oplus e_c \oplus e_r) + b^r \tag{3}$$

where $e_r$ is the tree embedding, $e_l$ is the embedding of the left subtree, $e_r$ is the embedding of the right subtree, $e_c$ is the connective embedding.

During the projection stage, we transform the tree embedding $e_r$ into the same representation space of LLMs through a projection layer, which includes two layers of neural networks:

$$e_t = W_2(W_1 e_r + b_1) + b_2 \tag{4}$$

where $W_1, W_2, b_1, b_2$ are the trainable parameters of the projection layer, $W_1 \in \mathbb{R}^{d'\times d}$, $W_2 \in \mathbb{R}^{d_{LLM}\times d'}$, $b_1 \in \mathbb{R}^{d'}$, $b_2 \in \mathbb{R}^{d_{LLM}}$, $d$ is dimension of hidden states in RoBERTa, $d'$ is the dimension of embedding space of the projection layer, $d_{LLM}$ is the dimension of embedding space of the target LLM. $e_t$ is the resulting tree-based soft prompt, which is then inserted into LLMs as a token representation within the input sequence.

### 4.3 Fallacy Training

The LLMs take the instruction prompt, textualized tree $h_t$, and tree-based soft prompt $e_t$ as input, and generate fallacy label as output. The loss is calculated between the generated text and golden label. The text embedding layer and self attention layers of LLMs are frozen. The tree-based soft prompt $e_t$ receives gradients and enables back propagation.

Table 2: The number of samples in train/dev/test set, the number of fallacy and no fallacy (benign) samples, and the number of fallacy types in each dataset.

| Dataset | Train | Dev | Test | Fallacy | Benign | Types |
|---|---|---|---|---|---|---|
| Argotario | 863 | 231 | 336 | 201 | 135 | 6 |
| Reddit | 2313 | 436 | 1849 | 1691 | 158 | 9 |
| Climate | 477 | 244 | 909 | 668 | 241 | 10 |
| Logic | 1625 | 206 | 436 | 436 | 0 | 13 |

# 5 Experiments

## 5.1 Datasets

We experiment with four datasets from various domains and genres. Table **??** shows their statistics.

- **Argotario** Habernal2017 collects fallacies from the general domain question-answering pairs. The dataset includes the following fallacy labels: Ad Hominem, Appeal to Emotion, Hasty Generalization, Irrelevant Authority, Red Herring, and No Fallacy. We use this dataset for both fallacy detection and classification experiments, and follow the dataset splitting method in Alhindi2022.
- **Reddit** Sahai2021 collects user generated posts from Reddit, and annotates logical fallacies into: Slippery Slope, Irrelevant Authority, Hasty Generalization, Black-and-White Fallacy, Ad Populum, Tradition Fallacy, Naturalistic Fallacy, Worse Problem Fallacy, and No Fallacy. This dataset is used for both fallacy detection and classification.
- **Climate** Alhindi2022 collects statements from articles in the climate change domain, and annotated the following fallacies: Evading the Burden of Proof, Cherry Picking, Red Herring, Strawman, Irrelevant Authority, Hasty Generalization, False Cause, False Analogy, Vagueness, and No Fallacy.
- **Logic** Jin2022 annotates logical fallacies in the educational materials into 13 types including Ad Hominem, Ad Populum, False Dilemma, False Cause, Circular Reasoning, Deductive Fallacy, Appeal to Emotion, Equivocation, Fallacy of Extension, Faulty Generalization, Intentional Fallacy, Fallacy of Credibility, Fallacy of Relevance. This dataset does not include No Fallacy class and is only used for fallacy classification.

## 5.2 Experimental Settings

To validate our approach, we experiment on two types of language models: a decoder-only model and an encoder-decoder model. For the decoder-only model, we choose the open-source large language model Llama-2 (llama-2-7b-chat-hf) Touvron2023. For the encoder-decoder model, we choose the Flan-T5-large model Chung2022. Both the models are trained in a generative setting, where they take the instruction and given text as input, and generate a fallacy label as output.

The fallacy detection task generates "Yes" or "No" label as output, while the fallacy classification task generates the name of each fallacy type. We follow Alhindi2022 to unify the different names of the same fallacy across datasets, such as False Dilemma is converted into Black-and-White Fallacy since they are the same fallacy. We also follow Alhindi2022 to feed the definitions of each fallacy type into the instruction prompt. The details of instruction prompt are explained in Appendix B. The maximum input length is set to be 1024, number of epochs is 10, weight decay is 1e-2, the gradient accumulation step is 4, learning rate for Llama-2 is 3e-4, and learning rate for Flan-T5 is 3e-5. The Llama-2 model is trained with LoRA Hu2021, with rank 8, alpha 16, dropout 0.05, and trainable modules include q_proj and v_proj.

## 5.3 Baselines

We compare our models with the baselines listed below. Besides the existing baselines, we also implement additional baselines based on the GPT and RoBERTa Liu2019 models:

- Sahai2021: a multi-granularity network is designed that trains sentence-level representation and the token-level representations jointly.
- Jin2022: a structure-aware framework is developed that forms a sequence-based logical pattern for each text by masking out the content words.
- Sourati2023b: a prototype-based reasoning method that injects background knowledge and explainable mechanisms into the language model.
- Sourati2023a: a case-based reasoning that retrieves similar cases from external sources based on goals, counterarguments, and explanation etc.
- Alhindi2022: a multi-task instruction tuning framework that learns the logical fallacies from multiple datasets collaboratively.
- **GPT-3.5**: we prompt the gpt-3.5-turbo model to automatically choose one of the fallacy labels for each text, and the prompt is listed in Appendix C.
- **GPT-3.5+Tree**: guide the gpt-3.5-turbo model to firstly reason the logical structure of each text, and then choose one of the fallacy labels through a chain-of-thought process Wei2023.
- **RoBERTa**: the RoBERTa model is used to encode the text and the average of word embedding is used as the text embedding. A classification head is built on top of the text embedding to classify labels.
- **RoBERTa+Tree**: we concatenate the text embedding with the logical structure tree embedding, and build classification head on top of the combined embedding to predict labels. The tree embedding is derived based on the method in Section **??**.

## 5.4 Fallacy Detection

The fallacy detection task identifies whether a given text contains logical fallacy or not, which is a binary classification task. The precision, recall, and F1 score of the fallacy class, as well as the micro F1 score (i.e., accuracy) are used as evaluation metrics. Table **??** presents the performance on the Argotario, Reddit, and Climate datasets.

The results demonstrate that incorporating the logical structure tree into Llama-2 and Flan-T5 models effectively improves both precision and recall for logical fallacy detection. This observation is consistent for both types of Llama-2 and Flan-T5 models across all the three datasets, which span various domains and genres. Compared to the baselines that lack logical structure information, our approach leads to the F1 score increased by up to 3.45%. This indicates that the logical structure tree is effective in capturing the difference in logical flows between fallacious and benign texts.

Moreover, informing the large language model GPT-3.5-turbo of logical structure information significantly improves fallacy detection under the zero-shot setting, resulting in a substantial improvement in the F1 score. This underscores the importance of infusing the logical structure information into LLMs for fallacy detection. Also, concatenating the logical structure tree embedding with the text embedding in the RoBERTa model also enhances the performance, which proves the usefulness of this logical structure tree embedding. Overall, incorporating the logical structure tree helps improve fallacy detection for various types of models.

## 5.5 Fallacy Classification

The fallacy classification task classifies the fallacy types for the fallacious text, which is a multi-class classification task excluding the No Fallacy class. The macro precision, recall, and F1 score, as well as the micro F1 score (i.e., accuracy) are used as evaluation metrics. Table **??** shows the results on the Argotario, Reddit, and Logic datasets.

The results demonstrate that integrating the logical structure tree into Llama-2 and Flan-T5 models notably enhances the performance of fallacy classification, with both precision and recall increased. This conclusion is valid across the three datasets from different domains and genres. Compared to the baselines without logical structure tree, our proposed approach significantly improves precision and recall, leading to an increase of up to 6.75% in the F1 score. This suggests that the logical structure tree effectively distinguishes the different logical patterns used in each fallacy type, and is applicable across various domains and genres.

Table 3: The results of logical fallacy detection on three datasets. The precision, recall, F1 score of fallacy class, and accuracy are reported. The rows "*+Tree" represent incorporating the logical structure tree into the model.

| | Argotario | | | | Reddit | | | | Climate | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Prec | Rec | F1 | Acc | Prec | Rec | F1 | Acc | Prec | Rec | F1 | Acc |
| Sahai et al. (2021) | 92.86 | 74.72 | 81.18 | 83.87 | 75.55 | 83.42 | 79.40 | 86.19 | 66.16 | 75.65 | 70.45 | 79.40 |
| GPT-3.5 | 74.61 | 75.55 | 75.08 | 75.38 | 69.57 | 54.17 | 60.61 | 65.00 | 76.02 | 69.20 | 72.45 | 73.88 |
| GPT-3.5+Tree | 81.91 | 84.37 | 83.52 | 86.02 | 85.08 | 89.50 | 87.19 | 88.40 | 83.47 | 86.86 | 85.01 | 87.19 |
| RoBERTa | 77.15 | 81.65 | 79.26 | 81.78 | 67.86 | 69.31 | 68.53 | 70.05 | 77.18 | 81.87 | 79.41 | 84.80 |
| RoBERTa+Tree | 72.48 | 75.07 | 73.57 | 76.72 | 69.85 | 72.24 | 70.96 | 73.73 | 68.50 | 69.17 | 68.80 | 69.17 |
| Flan-T5 | 94.56 | 100.00 | 97.11 | 100.00 | 79.45 | 81.78 | 80.45 | 81.78 | 66.16 | 69.17 | 67.52 | 69.17 |
| Flan-T5+Tree | 93.48 | 100.00 | 96.63 | 100.00 | 79.26 | 81.78 | 80.35 | 81.78 | 66.16 | 69.17 | 67.52 | 69.17 |
| Llama-2 | 83.98 | 86.19 | 85.01 | 88.40 | 83.15 | 85.71 | 84.37 | 87.19 | 77.90 | 80.52 | 79.13 | 82.40 |
| Llama-2+Tree | 83.52 | 85.25 | 84.34 | 86.02 | 83.47 | 86.72 | 85.01 | 87.19 | 77.19 | 83.95 | 80.45 | 84.80 |

In addition, our approach based on the logical structure tree outperforms the previous methods that may lack logical relations information. This highlights the necessity to infuse the logical relations into LLMs for fallacy classification. Besides, our approach achieves higher performance than the baselines that overlook content words. This indicates that analyzing content words also plays an essential role in fallacy reasoning. The logical structure tree connects the logical relations and content arguments together to form a cohesive logical structure, representing the hierarchical logical flow and thereby improving fallacy classification.

## 5.6 Ablation Study

The ablation study of the two designed strategies to incorporate the logical structure tree into LLMs is shown in Table **??**, where we take Llama-2 model as an example. The upper rows show the results of fallacy detection on the three datasets, and the lower rows show the results of fallacy classification.

The results demonstrate that both the textualized tree and tree-based soft prompt brings improvement for fallacy detection and classification across multiple datasets. This proves that the textualized tree and tree-based soft prompt are complementary with each other: the textualized tree enriches the instruction prompt with logical structure information, and the tree-based soft prompt enables direct learning from the tree embedding. Comparing across these two strategies, the soft prompt usually achieves better performance than the hard text prompt, and exhibits higher recall. Combining the two strategies together leads to the best performance, achieving the highest precision and recall.

## 5.7 Effect on Different Fallacy Types

We further analyze the F1 score change across each fallacy type in the fallacy classification task. The Llama-2 model is used as an example to show the performance change before and after incorporating the logical structure tree. Table **??** presents the F1 score change across each fallacy type on Argotario dataset. The performance change across each fallacy type on the Reddit and Logic dataset are shown in Table **??** and Table **??**. We observe that the logical structure tree brings bigger improvements for the fallacy types such as Red Herring, Hasty Generalization, Irrelevant Authority, Ad Populum, Extension Fallacy, Equivocation, Circular Reasoning etc. One possible explanation is that these fallacy types usually employ certain logical relations or logical patterns to persuade the readers. However, the performance

Table 4: The results of logical fallacy classification on three datasets. The macro precision, recall, F1 score, and accuracy are reported. The rows "*+Tree" represent incorporating the logical structure tree into the model.

| | Argotario | | | | Reddit | | | | Logic | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Prec | Rec | F1 | Acc | Prec | Rec | F1 | Acc | Prec | Rec | F1 | Acc |
| Jin et al. (2022) | 41.65 | 49.77 | 45.32 | 57.97 | 31.32 | 38.98 | 34.73 | 55.98 | 32.48 | 40.26 | 35.97 | 55.92 |
| Sourati et al. (2023b) | 59.51 | 59.59 | 59.55 | 62.72 | 55.91 | 49.22 | 52.34 | 57.90 | 58.41 | 49.81 | 53.78 | 57.96 |
| Sourati et al. (2023a) | 60.35 | 61.11 | 60.72 | 71.99 | 55.62 | 65.29 | 59.98 | 70.76 | 55.25 | 63.80 | 59.23 | 66.30 |
| Alhindi et al. (2022) | 59.67 | 55.98 | 57.76 | 60.03 | 38.14 | 36.93 | 37.52 | 62.50 | 63.67 | 63.10 | 63.38 | 66.40 |
| Sahai et al. (2021) | 32.58 | 40.59 | 36.15 | 59.66 | 47.67 | 63.10 | 54.30 | 64.88 | 42.28 | 47.99 | 44.95 | 64.88 |
| GPT-3.5 | 63.97 | 58.77 | 61.25 | 65.70 | 62.00 | 63.60 | 62.79 | 69.37 | 69.23 | 73.49 | 71.30 | 74.16 |
| GPT-3.5+Tree | 65.70 | 62.78 | 64.21 | 65.63 | 63.60 | 69.14 | 66.25 | 70.03 | 63.38 | 70.03 | 66.54 | 70.03 |
| RoBERTa | 58.46 | 62.95 | 60.62 | 63.92 | 76.10 | 81.34 | 78.63 | 84.84 | 76.01 | 81.29 | 78.57 | 83.95 |
| RoBERTa+Tree | 58.01 | 62.78 | 60.30 | 64.09 | 76.47 | 81.25 | 78.79 | 83.63 | 76.47 | 81.29 | 78.81 | 83.63 |
| Flan-T5 | 60.91 | 65.23 | 62.99 | 65.63 | 57.40 | 62.12 | 59.67 | 63.29 | 58.46 | 62.95 | 60.62 | 63.92 |
| Flan-T5+Tree | 57.40 | 62.12 | 59.67 | 63.29 | 58.01 | 62.78 | 60.30 | 64.09 | 58.01 | 62.78 | 60.30 | 64.09 |
| Llama-2 | 60.79 | 62.63 | 61.69 | 64.34 | 77.81 | 80.98 | 79.36 | 82.87 | 65.52 | 68.71 | 67.08 | 70.70 |
| Llama-2+Tree | 65.63 | 63.29 | 64.44 | 65.63 | 84.84 | 83.68 | 84.26 | 84.84 | 74.16 | 73.49 | 73.82 | 74.16 |

increase is less noticeable for the fallacy types such as Appeal to Emotion and Ad Hominem. It may be due to the reason that these fallacies rely more on the emotional or sentimental language instead of logical relations.

# 6 Limitations

We have compiled a set of connective words and phrases for the ten logical relations, as detailed in Table **??**. While we have included the common connectives in this set, it may not contain all the possible connectives. The logical structure tree that is constructed based on this connective words set demonstrates its usefulness in fallacy reasoning. Future work can be expanding this connectives set and investigating the effects of various connectives, so that we can better identify and mitigate them.

The release of code, datasets, and model should be used for mitigating logical fallacies, instead of expanding or disseminating the misinformation.

# 7 Conclusion

This paper detects and classifies fallacies. We propose a logical structure tree to explicitly represent and track the hierarchical logic flow among relation connectives and their arguments. We also design two strategies to incorporate this logical structure tree into LLMs for fallacy reasoning. Extensive experiments demonstrate the effectiveness of our approach based on the logical structure tree.

Table 5: The results of ablation study. The precision, recall, F1 score of fallacy class are reported for fallacy detection (upper rows). The macro precision, recall, F1 score are reported for fallacy classification (lower rows).

| Model | Prec | Rec | F1 |
|---|---|---|---|
| **Fallacy Detection** | | | |
| Argotario | | | |
| + textualized tree | 83.98 | 86.19 | 85.01 |
| + tree-based soft prompt | 83.15 | 85.71 | 84.37 |
| + both (full model) | 83.52 | 85.25 | 84.34 |
| Reddit | | | |
| + textualized tree | 83.47 | 86.72 | 85.01 |
| + tree-based soft prompt | 83.15 | 85.71 | 84.37 |
| + both (full model) | 83.47 | 86.72 | 85.01 |
| Climate | | | |
| + textualized tree | 77.90 | 80.52 | 79.13 |
| + tree-based soft prompt | 77.19 | 83.95 | 80.45 |
| + both (full model) | 77.19 | 83.95 | 80.45 |
| **Fallacy Classification** | | | |
| Argotario | | | |
| + textualized tree | 60.79 | 62.63 | 61.69 |
| + tree-based soft prompt | 64.34 | 65.63 | 64.98 |
| + both (full model) | 65.63 | 63.29 | 64.44 |
| Reddit | | | |
| + textualized tree | 77.81 | 80.98 | 79.36 |
| + tree-based soft prompt | 82.87 | 84.84 | 83.85 |
| + both (full model) | 84.84 | 83.68 | 84.26 |
| Logic | | | |
| + textualized tree | 65.52 | 68.71 | 67.08 |
| + tree-based soft prompt | 70.70 | 74.16 | 72.39 |
| + both (full model) | 74.16 | 73.49 | 73.82 |

## Ethical considerations

This paper aims to detect and classify logical fallacies. Logical fallacy is the error or flaws in the reasoning, and can occur in various human communications. Logical fallacies can lead to harmful consequences for society, such as spreading misinformation or introducing societal bias. The goal of this research is to understand logical fallacies, and help mitigate their harmful effects.

## Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from National Science Foundation via the award IIS2127746. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

## References

[Abd-Eldayem2023] Rehab Mohamed Ahmed Abd-Eldayem. 2023. The relationship between cognitive bias and logical fallacies in egyptian society. *Social Sciences*, 12(6):281–293.

[Aimeur et al.2023] Esma Aimeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

[Alhindi et al.2022] Tariq Alhindi, Tuhin Chakraborty, Elena Musi, and Smaranda Muresan. 2022. Multitask instruction-based prompting for fallacy recognition. In *Proceedings of the 2022 Conference on Empirical Meth-*

Table 6: The F1 score change across each fallacy type of fallacy classification on Argotario dataset. The fallacy types include Ad Hominem, Emotional Language, Hasty Generalization, Irrelevant Authority, and Red Herring.

| Fallacy Type | Llama-2 | Llama-2+Tree |
|---|---|---|
| Ad Hominem | 60.79 | 63.16 |
| Emotional Language | 67.33 | 72.16 |
| Hasty Generalization | 55.38 | 61.29 |
| Irrelevant Authority | 63.16 | 67.80 |
| Red Herring | 49.35 | 55.17 |
| Macro F1 | 59.20 | 63.92 |

Table 7: The F1 score change across each fallacy type of fallacy classification on Reddit dataset. The fallacy types include Slippery Slope, Irrelevant Authority, Hasty Generalization, Black-and-White Fallacy, Ad Populum, Tradition Fallacy, Naturalistic Fallacy, and Worse Problem Fallacy.

| Fallacy Type | Llama-2 | Llama-2+Tree |
|---|---|---|
| Slippery Slope | 86.96 | 88.89 |
| Irrelevant Authority | 82.05 | 92.31 |
| Hasty Generalization | 69.57 | 77.27 |
| Black-and-White Fallacy | 63.41 | 65.22 |
| Ad Populum | 68.29 | 82.93 |
| Tradition Fallacy | 81.82 | 87.18 |
| Naturalistic Fallacy | 90.00 | 95.25 |
| Worse Problem Fallacy | 75.56 | 82.61 |
| Macro F1 | 77.21 | 83.95 |

*ods in Natural Language Processing*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Armitage and Vaccari2021] Rachel Armitage and Cristian Vaccari. 2021. Misinformation and disinformation. In *The Routledge companion to media disinformation and populism*, pages 38–48. Routledge.

[Barclay2018] Donald A Barclay. 2018. *Fake news, propaganda, and plain old lies: How to find trustworthy information in the digital age*. Rowman & Littlefield.

[Beisecker et al.2024] Sven Beisecker, Christian Schlereth, and Sebastian Hein. 2024. Shades of fake news: How fallacies influence consumers' perception. *European Journal of Information Systems*, 33(1):41–60.

[Bonial et al.2022] Claire Bonial, Austin Blodgett, Taylor Hudson, Stephanie M. Lukin, Jeffrey Micher, Douglas Summers-Stay, Peter Sutor, and Clare Voss. 2022. The search for agreement on logical fallacy annotation of an infodemic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4430–4438, Marseille, France. European Language Resources Association.

[Chen et al.2021] Zeming Chen, Qiyue Gao, and Lawrence S. Moss. 2021. Neurallog: Natural language inference with joint neural and logical reasoning. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 78–88, Online. Association for Computational Linguistics.

[Chung et al.2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. Preprint, arXiv:2210.11416.

[Cotton2018] Christian Cotton. 2018. Argument from fallacy. *Bad arguments: 100 of the most important fallacies in Western philosophy*, pages 125–127.

Table 8: The F1 score change across each fallacy type of fallacy classification on Logic dataset. The fallacy types include Ad Hominem, Ad Populum, False Dilemma (Black-and-White Fallacy), False Cause, Circular Reasoning, Deductive Fallacy, Appeal to Emotion (Emotional Language), Equivocation, Fallacy of Extension, Faulty Generalization (Hasty Generalization), Intentional Fallacy, Fallacy of Credibility (Irrelevant Authority), Fallacy of Relevance (Red Herring).

| Fallacy Type | Llama-2 | Llama-2+Tree |
|---|---|---|
| Ad Hominem | 82.35 | 80.46 |
| Ad Populum | 72.41 | 87.50 |
| False Dilemma | 78.57 | 78.57 |
| False Cause | 68.42 | 66.67 |
| Circular Reasoning | 61.90 | 75.68 |
| Deductive Fallacy | 62.07 | 66.67 |
| Emotional Language | 66.67 | 65.22 |
| Equivocation | 25.00 | 44.44 |
| Extension Fallacy | 60.00 | 72.22 |
| Hasty Generalization | 78.13 | 81.03 |
| Intentional Fallacy | 34.48 | 38.71 |
| Irrelevant Authority | 64.71 | 68.97 |
| Red Herring | 65.00 | 78.05 |
| Macro F1 | 63.05 | 69.55 |


[Creswell et al.2022] Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. Preprint, arXiv:2205.09712.

[Da San Martino et al.2019] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. 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, Hong Kong, China. Association for Computational Linguistics.

[Da San Martino et al.2020] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

[Devatmane et al.2023] Nicolas Devatmane, Philippe Muller, and Chloé Braud. 2023. An integrated approach for political bias prediction and explanation based on discursive structure. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11196–11211, Toronto, Canada. Association for Computational Linguistics.

[Fantino et al.2003] Edmund Fantino, Stephanie Stolarz-Fantino, and Anton Navarro. 2003. Logical fallacies: A behavioral approach to reasoning. *The Behavior Analyst Today*, 4(1):109.

[Feng et al.2023] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

[Goffredo et al.2022] Pierpaolo Goffredo, Shokeh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4143–4149.

[Goffredo et al.2023] Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

[Greenwell et al.2006] William S Greenwell, John C Knight, C Michael Holloway, and Jacob J Pease. 2006. A taxonomy of fallacies in system safety arguments. In *24th International System Safety Conference*.

[Guess and Lyons2020] Andrew M Guess and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10.

[Habernal et al.2017] Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

[Habernal et al.2018] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *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. Association for Computational Linguistics.

[Hu et al.2021] 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. Preprint, arXiv:2106.09685.

[Jiao et al.2022] Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. MERIT: Meta-Path Guided Contrastive Learning for Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3496–3509, Dublin, Ireland. Association for Computational Linguistics.

[Jin et al.2022] Zheng Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Lei and Huang2022] Yuanyuan Lei and Ruihong Huang. 2022. Few-shot (dis)agreement identification in online discussions with regularized and augmented meta-learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5581–5593, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Lei and Huang2023a] Yuanyuan Lei and Ruihong Huang. 2023a. Discourse structures guided fine-grained propaganda identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.

[Lei and Huang2023b] Yuanyuan Lei and Ruihong Huang. 2023b. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822, Singapore. Association for Computational Linguistics.

[Lei and Huang2024] Yuanyuan Lei and Ruihong Huang. 2024. Sentence-level media bias analysis with event relation graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5225–5238, Mexico City, Mexico. Association for Computational Linguistics.

[Lei et al.2022] Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Lei et al.2024a] Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024a. EMONA: Event-level moral opinions in news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5239–5251, Mexico City, Mexico. Association for Computational Linguistics.

[Lei et al.2024b] Yuanyuan Lei, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Ruihong Huang, and Dong Yu. 2024b. Polarity calibration for opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5211–5224, Mexico City, Mexico. Association for Computational Linguistics.

[Li et al.2019] Quanzhi Li, Qiong Zhang, Luo Si, and Yingchi Liu. 2019. Rumor detection on social media: Datasets, methods and opportunities. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 66–75, Hong Kong, China. Association for Computational Linguistics.

[Li et al.2022] Yitian Li, Jidong Tian, Wenqing Chen, Caoyun Fan, Hao He, and Yaohui Jin. 2022. To what extent do natural language understanding datasets correlate to logical reasoning? a method for diagnosing logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1708–1717, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

[Lin et al.2020] Timothy PH Lin, Kelvin H Wan, Suber S Huang, Jost B Jonas, David SC Hui, and Dennis SC Lam. 2020. Death tolls of covid-19: where come the fallacies and ways to make them more accurate. *Global Public Health*, 15(10):1582–1587.

[Liu et al.2019] 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. Preprint, arXiv:1907.11692.

[Lundy2023] Morgan Lundy. 2023. Tiktok and covid-19 vaccine misinformation: New avenues for misinformation spread, popularinfodemic topics, and dangerous logical fallacies. *International Journal of Communication*, 17:24.

[Ma et al.2018] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

[Mancini et al.2024] Eleonora Mancini, Federico Ruggeri, and Paolo Torroni. 2024. Multimodal fallacy classification in political debates. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–178, St. Julian's, Malta. Association for Computational Linguistics.

[Musi and Reed2022] Elena Musi and Chris Reed. 2022. From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, 33(3):349–370.

[Musi et al.2022] Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies*, 12(3).

[Nakpih and Santini2020] Callistus Ireneous Nakpih and Simone Santini. 2020. Automated discovery of logical fallacies in legal argumentation. *International Journal of Artificial Intelligence and Applications (IJNA)*, 11.

[Olausson et al.2023] Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

[Oshikawa et al.2020] Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

[Pan et al.2023] Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

[Parmar et al.2024] Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13679–13707, Bangkok, Thailand. Association for Computational Linguistics.

[Pauli et al.2022] Amalie Pauli, Leon Derczynski, and Ira Assent. 2022. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 89–100, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

[Pi et al.2022] Xinyu Pi, Wanjun Zhong, Yan Gao, Nan Duan, and Jian-Guang Lou. 2022. Logigan: Learning logical reasoning via adversarial pre-training. Preprint, arXiv:2205.08794.

[Prasad et al.2008] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[Qi et al.2020] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[Rashkin et al.2017] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

[Risen et al.2007] Jane Risen, Thomas Gilovich, R Stemberg, D Halpern, and H Roediger. 2007. Informal logical fallacies. *Critical thinking in psychology*, 110.

[Sahai et al.2021] Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. 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 644–657, Online. Association for Computational Linguistics.

[Sanyal et al.2023] Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu, and Xiang Ren. 2023. APOLLO: A simple approach for adaptive pretraining of language models for logical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6308–6321, Toronto, Canada. Association for Computational Linguistics.

[Sheng et al.2021] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

[Sourati et al.2023a] Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-An Sandlin, and Alain Mermoud. 2023a. Case-based reasoning with language models for classification of logical fallacies. Preprint, arXiv:2301.11879.

[Sourati et al.2023b] Zhivar Sourati, Filip Ilievski, Hông-An Sandlin, and Alain Mermoud. 2023b. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

[Stab and Gurevych2017] Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

[Tindale2007] Christopher W Tindale. 2007. *Fallacies and argument appraisal*. Cambridge University Press.

[Touvron et al.2023] 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.

[Walton1987] Douglas N Walton. 1987. What is a fallacy. In Frans H. van Eemeren, Rob Grootendorst, J. Anthony Blair, Charles A. Willard (eds.): *Argumentation: Across the Lines of Discipline*. Dordrecht/Providence, Foris Publications, pages 323–330.

[Walton2010] Douglas Walton. 2010. Why fallacies appear to be better arguments than they are. *Informal Logic*, 30(2):159–184.

[Walton et al.2008] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

[Wei et al.2023] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. Preprint, arXiv:2201.11903.

[Xu et al.2023] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. Preprint, arXiv:2306.09841.

[Yang et al.2017] Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989, Copenhagen, Denmark. Association for Computational Linguistics.

[Zhang et al.2023] Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. 2023. Improved logical reasoning of language models via differentiable symbolic programming. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3062–3077, Toronto, Canada. Association for Computational Linguistics.

[Zhao et al.2023] Hongyu Zhao, Kangrui Wang, Mo Yu, and Hongyuan Mei. 2023. Explicit planning helps language models in logical reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11155–11173, Singapore. Association for Computational Linguistics.

[Zhou et al.2023] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. Preprint, arXiv:2205.10625.

# A Statistical Analysis of Logical Relations

Table **??** presents the ratio of samples that contain the ten logical relations in fallacy and no fallacy classes, where we take the Argotario Habernal2017 and Reddit Sahai2021 datasets as examples. Further, Table **??** shows the ratio of samples that contain the ten logical relations in each fallacy type, where we take the Logic dataset Jin2022 as an example.

Table 9: The ratio (%) of samples that contain the ten logical relations in fallacy and no fallacy classes in the development set of Argotario (the first two rows) and Reddit (the latter two rows) datasets.

| Dataset | conj | alt | rest | inst | cont | conc | anal | temp | cond | caus |
|---|---|---|---|---|---|---|---|---|---|---|
| Argotario fallacy | 37.96 | 28.13 | 46.72 | 40.63 | 1.46 | 3.13 | 0.73 | 0.00 | 48.91 | 42.19 |
| Argotario no fallacy | 1.46 | 3.13 | 6.57 | 1.56 | 10.95 | 7.81 | 16.06 | 15.63 | 69.34 | 56.25 |
| Reddit fallacy | 64.04 | 50.31 | 75.44 | 69.63 | 4.39 | 3.71 | 2.92 | 1.53 | 61.54 | 61.18 |
| Reddit no fallacy | 8.19 | 7.98 | 16.67 | 19.94 | 26.90 | 25.46 | 34.80 | 33.44 | 79.24 | 73.01 |

Table 10: The ratio (%) of samples that contain the ten logical relations in each fallacy type in the Logic dataset. The fallacy types include Ad Hominem, Ad Populum, False Dilemma (Black-and-White Fallacy), False Cause, Circular Reasoning, Deductive Fallacy, Appeal to Emotion (Emotional Language), Equivocation, Fallacy of Extension, Faulty Generalization (Hasty Generalization), Intentional Fallacy, Fallacy of Credibility (Irrelevant Authority), Fallacy of Relevance (Red Herring).

| Fallacy Type | conj | alt | rest | inst | cont | conc | anal | temp | cond | caus |
|---|---|---|---|---|---|---|---|---|---|---|
| Ad Hominem | 30.22 | 20.88 | 18.34 | 46.74 | 24.24 | 28.09 | 41.86 | 42.10 | 54.71 | 38.99 |
| Ad Populum | 29.46 | 39.25 | 35.96 | 60.44 | 47.46 | 79.81 | 62.72 | 38.63 | 63.63 | 59.68 |
| False Dilemma | 71.05 | 73.58 | 52.83 | 48.21 | 66.35 | 67.54 | 0.44 | 0.63 | 0.91 | 0.00 |
| False Cause | 0.00 | 0.00 | 0.00 | 2.32 | 0.00 | 0.00 | 0.00 | 1.88 | 0.63 | 0.89 |
| Circular Reasoning | 4.67 | 0.00 | 64.44 | 27.21 | 36.69 | 36.68 | 39.61 | 50.38 | 63.16 | 62.26 |
| Deductive Fallacy | 39.93 | 60.71 | 41.12 | 55.26 | 2.22 | 1.89 | 2.75 | 1.18 | 0.00 | 0.82 |
| Emotional Language | 1.55 | 7.89 | 0.94 | 1.88 | 4.46 | 2.80 | 0.00 | 7.55 | 5.06 | 1.83 |
| Equivocation | 3.55 | 3.03 | 19.83 | 7.75 | 5.26 | 11.32 | 8.49 | 5.36 | 3.73 | 4.38 |
| Extension Fallacy | 4.38 | 12.00 | 10.76 | 15.59 | 37.87 | 11.36 | 17.35 | 18.60 | 31.57 | 11.32 |
| Hasty Generalization | 23.27 | 18.75 | 7.47 | 20.17 | 12.44 | 10.12 | 28.44 | 11.24 | 9.09 | 10.79 |
| Intentional Fallacy | 28.68 | 28.94 | 18.86 | 31.13 | 25.00 | 16.82 | 12.28 | 76.89 | 72.15 | 50.45 |
| Irrelevant Authority | 86.98 | 83.33 | 76.03 | 63.56 | 76.31 | 87.73 | 69.49 | 67.85 | 84.11 | 74.56 |
| Red Herring | 74.56 | 34.49 | 58.97 | 0.87 | 0.82 | 45.97 | 1.85 | 6.91 | 18.22 | 19.75 |
| Overall | 74.37 | 34.49 | 58.97 | 0.87 | 0.82 | 45.97 | 1.85 | 6.91 | 18.22 | 19.75 |

# B Instruction Prompts

## B.1 Prompt for Fallacy Detection

The instruction prompt for the Llama-2 or Flan-T5 baseline model is: "The task is to detect whether the Text contains logical fallacy or not. The logical fallacy can be [fallacy name (fallacy definition)]. Please answer Yes if the Text contains logical fallacy, else answer No. Text: [text]. Answer:"

The instruction prompt that incorporates the textualized tree into the Llama-2 or Flan-T5 model is: "The task is to detect whether the Text contains logical fallacy or not. The logical fallacy can be [fallacy name (fallacy definition)]. The logical relations in the Text are presented in this table: argument 1, logical relation, argument 2 [textualized tree]. Please answer Yes if the Text contains logical fallacy, else answer No. Text: [text]. Answer:"

## B.2 Prompt for Fallacy Classification

The instruction prompt for the Llama-2 or Flan-T5 baseline model is: "The task is to classify the fallacy type of the Text. Choose one answer from these fallacy types: [fallacy names list]. The definitions of each fallacy type are as follows. [fallacy name: fallacy definition]. Please classify the fallacy type of the Text. Text: [text]. Answer:"

The instruction prompt that incorporates the textualized tree into the Llama-2 or Flan-T5 model is: "The task is to classify the fallacy type of the Text. Choose one answer from these fallacy types: [fallacy names list]. The definitions of each fallacy type are as follows. [fallacy name: fallacy definition]. The logical relations in the Text are presented in this table: argument 1, logical relation, argument 2 [textualized tree]. Please classify the fallacy type of the Text. Text: [text]. Answer:"

# C Fallacy Definitions

## C.1 Argotario Dataset

The Argotario dataset Habernal2017 includes five fallacy types: Ad Hominem, Appeal to Emotion, Hasty Generalization, Irrelevant Authority, Red Herring. The name of Appeal to Emotion is converted into Emotional Language. The definitions of these fallacy types which are used in the instruction prompt are:

- Ad Hominem: the text attack a person instead of arguing against the claims.
- Emotional Language: the text arouse non-rational emotions.
- Hasty Generalization: the text draw a broad conclusion based on a limited sample of population.
- Irrelevant Authority: the text cite an authority but the authority lacks relevant expertise.
- Red Herring: the text diverge the attention to irrelevant issues.

## C.2 Reddit Dataset

The Reddit dataset Sahai2021 includes eight fallacy types and their label names are: Slippery Slope, Irrelevant Authority, Hasty Generalization, Black-and-White Fallacy, Ad Populum, Tradition Fallacy, Naturalistic Fallacy, Worse Problem Fallacy. The definitions of these fallacy types which are used in the instruction prompt are:

- Slippery Slope: the text suggest taking a small initial step leads to a chain of related events culminating in significant effect.
- Irrelevant Authority: the text cite an authority but the authority lacks relevant expertise.
- Hasty Generalization: the text draw a broad conclusion based on a limited sample of population.
- Black-and-White Fallacy: the text present two alternative options as the only possibilities.
- Ad Populum: the text affirm something is true because the majority thinks so.
- Tradition Fallacy: the text argue the action has always been done in the tradition.
- Naturalistic Fallacy: the text claim something is good or bad because it is natural or unnatural.
- Worse Problem Fallacy: the text justify an issue by arguing more severe issues exists.

## C.3 Climate Dataset

The Climate dataset Alhindi2022 includes the following fallacy types: Evading Burden of Proof, Cherry Picking, Red Herring, Strawman, False Authority, Hasty Generalization, False Cause, Post Hoc, False Analogy, Vagueness. The name of False Authority is replaced by Irrelevant Authority. The class of Post Hoc is combined into False Cause. The definitions of these fallacy types which are used in the instruction prompt are:

- Evading Burden of Proof: the text make a claim without evidence or supporting argument.
- Cherry Picking: the text selectively present partial evidence to support a claim.
- Red Herring: the text diverge the attention to irrelevant issues.
- Strawman: the text distort the claim to another one to make it easier to attack.
- Irrelevant Authority: the text cite an authority but the authority lacks relevant expertise.
- Hasty Generalization: the text draw a broad conclusion based on a limited sample of population.
- False Cause: the text assume two correlated events must also have a causal relation.
- False Analogy: the text assume two alike things must be alike in other aspects.
- Vagueness: the text use ambiguous words, terms, or phrases.

## C.4 Logic Dataset

The Logic dataset Jin2022 annotates 13 types of fallacy: Ad Hominem, Ad Populum, False Dilemma (Black-and-White Fallacy), False Cause, Circular Reasoning, Fallacy of Logic (Deductive Fallacy), Appeal to Emotion (Emotional Language), Equivocation, Fallacy of Extension (Extension Fallacy), Faulty Generalization (Hasty Generalization), Intentional Fallacy, Fallacy of Credibility (Irrelevant Authority), Fallacy of Relevance (Red Herring). The names in the parenthesis are the replaced names used in the instruction prompt. The definitions of these fallacy types which are used in the instruction prompt are:

- Ad Hominem: the text attack a person instead of arguing against the claims.
- Ad Populum: the text affirm something is true because the majority thinks so.
- Black-and-White Fallacy: the text present two alternative options as the only possibilities.
- False Cause: the text assume two correlated events must also have a causal relation.
- Circular Reasoning: the end of the text come back to the beginning without having proven itself.
- Deductive Fallacy: the text has an error in the logical reasoning.
- Emotional Language: the text arouse non-rational emotions.
- Equivocation: the text use a key term in multiple senses, leading to ambiguous conclusions.
- Extension Fallacy: the text attack an exaggerated version of the opponent's claim.
- Hasty Generalization: the text draw a broad conclusion based on a limited sample of population.
- Intentional Fallacy: the text show intentional action to incorrectly support an argument.
- Irrelevant Authority: the text cite an authority but the authority lacks relevant expertise.
- Red Herring: the text diverge the attention to irrelevant issues.