Logical Fallacy Detection

Zhijing Jin[1,2,*] **Abhinav Lalwani**[3,*,†] **Tejas Vaidhya**[4,†] **Xiaoyu Shen**[5] **Yiwen Ding**[6]
Zhiheng Lyu[7,†] **Mrinmaya Sachan**[2] **Rada Mihalcea**[6] and **Bernhard Schölkopf**[1,2]
[1]Max Planck Institute, [2]ETH Zürich, [3]BITS Pilani, [4]IIT Kharagpur,
[5]Saarland Informatics Campus, [6]University of Michigan, [7]University of Hong Kong
`jinzhi@ethz.ch` `abhinav.lalwani@gmail.com`

Abstract

Reasoning is central to human intelligence. However, fallacious arguments are common, and some exacerbate problems such as spreading misinformation about climate change. In this paper, we propose the task of *logical fallacy detection*, and provide a new dataset (**LOGIC**) of logical fallacies generally found in text, together with an additional challenge set for detecting logical fallacies in climate change claims (**LOGICCLIMATE**). Detecting logical fallacies is a hard problem as the model must understand the underlying logical structure of the argument. We find that existing pretrained large language models perform poorly on this task. In contrast, we show that a simple structure-aware classifier outperforms the best language model by 5.46% F_1 scores on LOGIC and 4.51% on LOGICCLIMATE. We encourage future work to explore this task since (a) it can serve as a new reasoning challenge for language models, and (b) it can have potential applications in tackling the spread of misinformation.[1]

1 Introduction

Reasoning is the process of using existing knowledge to make inferences, create explanations, and generally assess things rationally by using logic (Aristotle, 1991). Human reasoning is, however, often marred with logical fallacies. Fallacious reasoning leads to disagreements, conflicts, endless debates, and a lack of consensus. In daily life, fallacious arguments can be as harmless as "*All tall people like cheese*" (faulty generalization) or "*She is the best because she is better than anyone else*" (circular claim). However, logical fallacies are also intentionally used to spread misinformation, for instance "*Today is so cold, so I don't believe in global warming*" (faulty generalization) or "*Global warming doesn't exist because the earth is not getting warmer*" (circular claim).

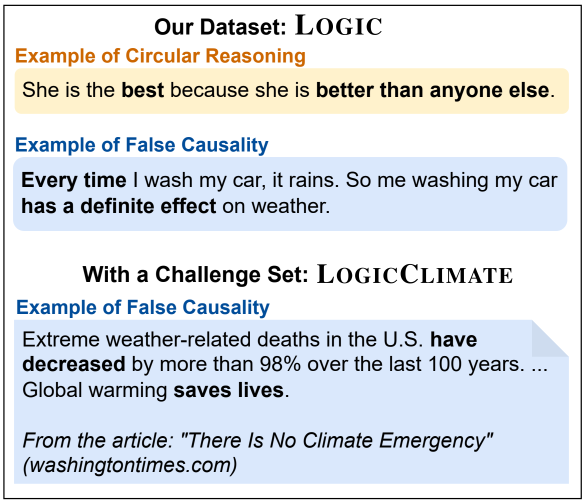

Figure 1: Our dataset consists of general logical fallacies (LOGIC) and an additional test set of logical fallacies in climate claims (LOGICCLIMATE).

In order to detect such fallacious arguments, we propose the task of *logical fallacy detection*. Logical fallacy detection methods can be helpful to tackle important social problems. For instance, these methods can be combined with fact-checkers (Riedel et al., 2017; Thorne et al., 2018) for misinformation detection as many claims can be factually correct but still fallacious. However, logical fallacy detection is challenging as it requires a model to discover egregious patterns of reasoning (Johnson and Blair, 2006; Damer, 2009).

To address this pressing need and encourage more work to detect reasoning flaws, we construct a dataset of logical fallacies consisting of general logical fallacies (**LOGIC**), and a challenging extrapolation set of climate claims (**LOGICCLIMATE**), as shown in Figure 1. We find that this task is challenging for 12 pretrained large language models, whose performances range from 8.62% to 53.31% micro F_1 scores on the LOGIC dataset.

[*] Equal contribution.
[†] Done during the research internship at ETH Zürich.
[1]Our dataset and code are available at https://github.com/causalNLP/logical-fallacy.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7180–7198
December 7-11, 2022 ©2022 Association for Computational Linguistics

Logical Fallacy	Examples
Faulty Generalization (18.01%)	"I met **a tall man** who loved to eat cheese. Now I believe that **all tall people** like cheese." "**Sometimes** flu vaccines don't work; therefore **vaccines are useless**."
Ad Hominem (12.33%)	"What can our new math teacher know? Have you seen **how fat** she is?" "I cannot listen to anyone **who does not share my social and political values**."
Ad Populum (9.47%)	"Everyone should like coffee: **95%** of teachers do!" "Killing thousands of people as a result of drug war campaign is **not a crime** to humanity because **millions of Filipino support it**."
False Causality (8.82%)	"**Every time** I wash my car, it rains. Me washing my car has **a definite effect** on the weather." "Every severe recession **follows** a Republican Presidency; therefore Republicans **are the cause of** recessions."
Circular Claim (6.98%)	"J.K. Rowling is a **wonderful** writer because she **writes so well**." "She is the **best** candidate for president because she **is better than** the other candidates!"
Appeal to Emotion (6.82%)	"It is an **outrage** that the school wants to remove the vending machines. This is **taking our freedom away**!" "Vaccines are so **unnatural**; it's **disgusting** that people are willing to put something like that in their body."
Fallacy of Relevance (6.61%)	"**Why are you worried** about poverty? **Look how many** children we abort every day." "**Why should we be worrying** about how the government treats Native people, **when** people in our city can't get a job"
Deductive Fallacy (6.21%)	"**It is possible** to fake the moon landing through special effects. **Therefore**, the moon landing was a fake using special effects." "Guns **are like** hammers—they're both tools with metal parts that could be used to kill someone. And **yet it would be ridiculous** to restrict the purchase of hammers, **so** restrictions on purchasing guns are **equally ridiculous**."
Intentional Fallacy (5.84%)	"**No one has** ever been able to prove that extraterrestrials exist, **so they must** not be real." "**It's common sense** that if you smack your children, they will stop the bad behavior. **So don't tell me** not to hit my kids."
Fallacy of Extension (5.76%)	"Their support of the discussion of sexual orientation issues is dangerous: **they advocate for** the exposure of children to sexually explicit materials, **which is wrong**." "They say we should cut back the defense budget. **Their position is that** they want to leave our nation completely defenseless!"
False Dilemma (5.76%)	"You're **either** for the war **or** against the troops." " **I don't want to** give up my car, so **I don't think** I can support fighting climate change."
Fallacy of Credibility (5.39%)	"My **professor**, who has **a Ph.D. in Astronomy**, once **told me** that ghosts are real. **Therefore, ghosts are real.**" "My **minister** says the Covid vaccine will cause genetic mutations. He **has a college degree**, and is a **holy man, so he must be right**."
Equivocation (2.00%)	"I don't see how you can say you're an **ethical person**. It's so hard to get you to do anything; your **work ethic** is so bad" "It is **immoral to kill an innocent human** being. Fetuses are innocent human beings. Therefore, it is **immoral to kill fetuses**."

Table 1: Examples of the 13 logical fallacy types and their distribution in the LOGIC dataset. To illustrate of the potential impact of learning logical fallacies, we select some examples with **neutral** impact that we manually identify, and some with **potentially negative** impact.

By analyzing our collected dataset, we identify that logical fallacies often rely on certain false patterns of reasoning. For example, a typical pattern in *false causality* in Figure 1 is "α co-occurs with $\beta \Rightarrow \alpha$ causes β." Motivated by this, we develop an approach to encourage language models to identify these underlying patterns behind the fallacies. In particular, we design a **structure-aware model** which identifies text spans that are semantically similar to each other, masks them out, and then feeds the masked text instances to a classifier. This structure distillation process can be implemented atop any pretrained language model. Experiments show that our model outperforms the best pretrained language model by 5.46% on LOGIC, and 4.51% on LOGICCLIMATE.

In summary, this paper makes the following contributions:

1. We propose a new task of logical fallacy classification.
2. We collect a dataset of 2,449 samples of 13 logical fallacy types, with an additional challenge set of 1,109 climate change claims with logical fallacies.
3. We conduct extensive experiments using 12 existing language models and show that these models have very limited performance on detecting logical fallacies.
4. We design a structure-aware classifier as a baseline model for this task, which outperforms the best language model.

5. We encourage future work to explore this task and enable NLP models to discover erroneous patterns of reasoning.

2 Logical Fallacy Dataset

First, we introduce our data. Our logical fallacy dataset consists of two parts: a) a set of common logical fallacies (LOGIC), and b) an additional challenge set of logically fallacious claims about climate change (LOGICCLIMATE).

2.1 Common Logical Fallacies: LOGIC

Data Collection The LOGIC dataset consists of common logical fallacy examples collected from various online educational materials meant to teach or test the understanding of logical fallacies among students. We automatically crawled examples of logical fallacies from three student quiz websites, Quizziz, study.com and ProProfs (resulting in around 1.7K samples), and manually collected fallacy examples from some additional websites recommended by Google search (resulting in around 600 samples). More data collection and filtering details are in Appendix A.2.

	# Samples	# Sents	# Tokens	Vocab
Total Data	2,449	4,934	71,060	7,624
Train	1,849	3,687	53,475	6,634
Dev	300	638	8,690	2,128
Test	300	609	8,895	2,184

Table 2: Statistics of the LOGIC dataset.

The entire LOGIC dataset contains 2,449 logical fallacy instances across 13 logical fallacy types. We randomly split the data into train, dev, and test sets; dataset statistics are shown in Table 2, and the distribution and examples of each type in Table 1. More details of each fallacy type are in Appendix A.3.

Comparison with Existing Datasets. Due to the challenges of data collection, all previous existing datasets on argument quality are of limited size. In Table 3, we draw a comparison among our dataset and two existing datasets: an argument sufficiency classification dataset (Stab and Gurevych, 2017), which proposes a binary classification task to identify whether the evidence can sufficiently support an argument, and another dataset dedicated for a specific type of logical fallacy called *ad hominem*, or *name-calling* (Habernal et al., 2018b) where the arguer attacks the person instead of the claim.

Dataset	# Claims	# Classes	Purpose
Arg. Suff.	1,029	Binary	Detect insufficiency
Ad Homi.	2,085	Binary	Detect name calling
LOGIC	**2,449**	**Multiple**	Detect **all** fallacy types

Table 3: Comparison of our logical fallacy dataset with two existing datasets, argument sufficiency classification (Stab and Gurevych, 2017) and ad hominem classification (Habernal et al., 2018b).

Compared to the existing datasets, our dataset has two advantages: (1) we have a larger number of claims in our dataset, and (2) our task serves the more general purpose of detecting all fallacy types instead of a single fallacy type. These two characteristics make our dataset significantly more challenging.

2.2 Challenge Set: LOGICCLIMATE

Logical fallacy detection on climate change is a small step towards promoting consensus and joint efforts to fight climate change. We are interested in whether models learned on the LOGIC dataset can generalize well to real-world discussions on climate change. Hence, we collect an extrapolation set LOGICCLIMATE which consists of all climate change news articles from the Climate Feedback website[2] by October 2021.

For each news article, we ask two different annotators who are native English speakers to go through each sentence in the article, and label all logical fallacies if applicable. Since directly classifying the logical fallacies at the article level is too challenging, we let the annotators select the text span while labeling the logical fallacies, and we compose each sample using the sentence containing the selected text span as logical fallacies. Details of the annotation process are described in Appendix A.4.

In total, the LOGICCLIMATE dataset has 1,079 samples of logical fallacies with on average 35.98 tokens per sample, and a vocabulary of 5.8K words. The label distributions are in Table 4. We provide examples of each fallacy in LOGICCLIMATE in Appendix A.5.

3 A Structure-Aware Model

The task of logical fallacy classification is unique in that logical fallacies are not just about the content words (such as the sentiment-carrying words in a

[2] https://climatefeedback.org/feedbacks/

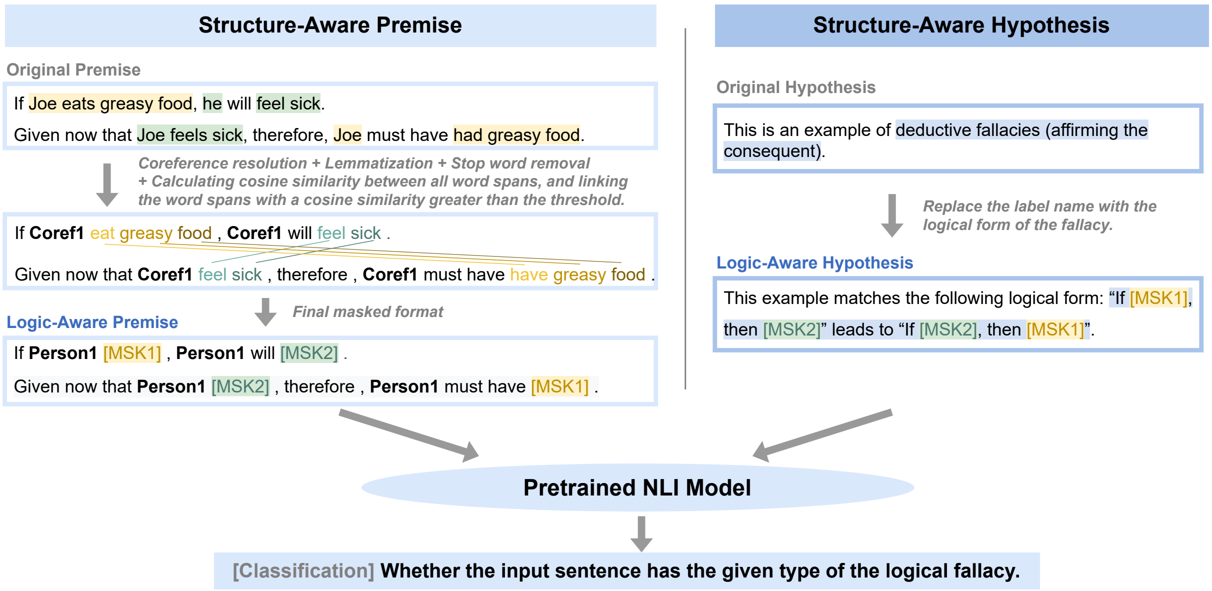

Figure 2: Our baseline model is a structure-aware classifier based on pretrained NLI model, with a structure-aware premise and structure-aware hypothesis. The structure-aware premise masks the content words to distill the argument structure. Specifically, we first resolve the coreferences, and then apply Sentence-BERT to match the lemmatized word spans (excluding the stopwords) whose contextualized embeddings have a cosine similarity larger than a certain threshold. And the structure-aware hypothesis uses the standard logical form of the given fallacy type.

Logical Fallacy Type	Frequency in Data
Intentional Fallacy	25.58%
Appeal to Emotion	11.37%
Faulty Generalization	10.18%
Fallacy of Credibility	9.90%
Ad Hominem	7.84%
Fallacy of Relevance	7.80%
Deductive Fallacy	6.50%
False Causality	5.11%
Fallacy of Extension	4.91%
Ad Populum	4.55%
False Dilemma	3.80%
Equivocation	1.94%
Circular Claim	0.51%

Table 4: Logical fallacy types and their frequencies in the LOGICCLIMATE dataset.

sentiment classification task), but more about the "form" or "structure" of the argument.

To advance the ability of models to detect fallacious logical structures, we draw inspirations from the history of logic (Russell, 2013). If we look into the time when Aristotle made his attempt to formulate a systematic study of logical, one of the most notable advancements is to move from contents to symbols, based on which Aristotle developed a system of rules (Gabbay and Woods, 2004). For example, he uses α, β, γ to distill arguments such as "Socrates is a man. All men are mortal. Therefore Socrates is mortal." into forms such as "α is a β. All β are γ. Therefore, α is γ.", where the variables act as placeholders. After establishing a system of valid and invalid argument structures, philosophers can refute a fallacious argument by comparing it to a list of fallacious logical forms (Aristotle, 2006).

Based on such inspirations, we propose a structure-aware classifier as a baseline model for our logical fallacy detection task. We first introduce a commonly used classification framework using pretrained models on natural language inference (NLI) in Section 3.1, and then we will propose our structure distillation process in Section 3.2.

3.1 Backbone: NLI-Based Classification with Pretrained Models

Motivated by the success of adapting NLI for classification tasks with unseen labels (Yin et al., 2019), we choose pretrained language models on NLI as the backbone of our logical fallacy classifier.

Specifically, a standard NLI-based pretrained language model for classification takes the sentence to classify as the *premise*. Then the model composes a *hypothesis* using the template of "This example is `[label name]`." The classifier checks whether the premise can entail the hypothesis. This NLI framework makes it easy for pretrained language models to adapt to unseen class labels such as our logical fallacy types.

3.2 Distilling Structure from Content

To build a model that encourages more attention to the *structure* of the text, we modify the premise and the hypothesis provided to the backbone NLI model (as shown in Figure 2): called the *structure-aware premise* and *structure-aware hypothesis*.

Structure-Aware Premise. Inspired by the process how ancient Greek philosophers refute an argument they have heard, we design an argument structure distiller by masking out content words in the premise (i.e., input text) and outputting a logical form with placeholders. In the example in Figure 2, "Jack is *a* good athlete. Jack comes from Canada. Therefore, *all* Canadians are good athletes.", we want the model to pay more attention to the structure as opposed to contents such as "good athletes." Thus, we build a distilled argument with placeholders "[MSK1] is a [MSK2]. [MSK1] comes from [MSK3]. Therefore, all [MSK3] are [MSK2]."

As shown in Figure 2, to distill the premise into the logical form, we identify all text spans that are paraphrases of each other and replace them with the same mask. Specifically, we first conduct coreference resolution using the `CoreNLP` package (Manning et al., 2014). Then, to identify word spans that are paraphrases of each other, we consider only non-stop words, lemmatize them via the `Stanza` package (Qi et al., 2020), and represent each word by its contextualized embedding generated by Sentence-BERT (Reimers and Gurevych, 2019), and calculate pair-wise cosine similarity. When the cosine similarity is larger than a threshold (by a grid search on the dev set), we identify the two words as similar. For illustration, we create a link between similar word pairs in Figure 2. When there are contiguous sequences of words that are linked to each other (e.g., "good athlete" and "good athletes"), we merge them and end up with two multi-word spans that are similar to each other. For each group i of similar text spans, we replace them with a mask token `[MSK`$_i$`]`.

Structure-Aware Hypothesis. NLI-based classification models (Yin et al., 2019) typically compose the hypothesis as a template sentence "This is an example of `[label name]`." However, in order to help our model perform a structure-aware matching of the logical fallacy instance, we also augment the hypothesis with the logical form for the logical fallacy type. For example, the logical form for *faulty generalization* in the example in Figure 2 is changed to: "[MSK1] has attribute [MSK2]. [MSK1] is a subset of [MSK3]. Therefore, all [MSK3] has attribute [MSK2]."

To look up the logical form of each fallacy, we refer to websites that introduce the logical fallacies, extract the expressions such as "Circular reasoning is often of the form: 'A is true because B is true; B is true because A is true."', and compile the logical forms using our masking format. We provide the list of logical forms in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Evaluation Metrics. Since the nature of the logical fallacy detection task is a multi-label classification with class imbalance, we use *micro* F_1 as the main evaluation metric. Additionally, we also report precision, recall and accuracy.

Baselines. We test the performance of 12 existing large language models, including five zero-shot models and seven finetuned models. For zero-shot models, we use the zero-shot classifier by `transformers` Python package (Wolf et al., 2020) implemented using RoBERTa large (Liu et al., 2019) and BART large (Lewis et al., 2020) finetuned on the multi-genre natural language inference (MNLI) task (Williams et al., 2018). We also include the task-aware representation of sentences (TARS) (Halder et al., 2020a) provided by FLAIR (Akbik et al., 2019). Moreover, we also try directly using GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020). For GPT-3, we designed a prompt for the auto-completion function to predict the label of text, and for GPT-2, we calculate the perplexity of every possible label with the text and choose the label with the lowest perplexity. See Appendix B.1 for more implementation details.

For finetuned baselines, we finetune seven commonly used pretrained language models on the LOGIC dataset, including ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), BigBird (Zaheer et al., 2020), DeBERTa (He et al., 2021), DistilBERT (Sanh et al., 2019), Electra (Clark et al., 2020), MobileBERT (Sun et al., 2020), and RoBERTa (Liu et al., 2019). See Appendix B.2 for implementation details.

Implementation Details. We describe the implementation details of our structure-aware classifier in Appendix B.3.

	F_1	P	R	Acc
Random	12.02	7.24	35.00	0.00
***Zero-shot classifiers directly tested on* LOGIC**				
TARS	8.62	3.86	6.67	2.33
BART-MNLI	11.05	6.63	33.67	0.00
GPT3	12.20	12.00	12.00	12.00
RoBERTa-MNLI	12.22	7.51	36.00	0.33
GPT2	**13.67**	13.67	13.67	13.67
***Finetuned and tested on* LOGIC**				
ALBERT	12.50	6.67	100.00	0.00
BigBird	15.02	8.61	90.00	0.33
DistilBERT	26.96	22.06	74.00	4.67
MobileBERT	35.68	29.05	71.00	7.33
BERT	45.80	40.73	73.67	18.00
DeBERTa	50.29	45.79	73.00	24.67
Electra	53.31	51.59	72.33	35.66
Electra-*StructAware*	**58.77**	55.25	63.67	47.67
Ablation study on the proposed model				
Raw Prem. + Str. Hypo.	56.72	54.87	76.67	37.67
Str. Prem. + Raw Hypo.	44.56	39.74	71.00	18.33

Table 5: Model performance on LOGIC by the ascending order of the main metric, micro F_1 (F_1). In addition, we also report the precision (P), recall (R), and accuracy (Acc). In the ablation study, we report the performance of two settings: a raw premise (i.e., keeping the original text input) with a structure-aware hypothesis (Raw Prem. + Str. Hypo.), and a structure-aware premise with a raw hypothesis (Str. Prem. + Raw Hypo.).

4.2 Main Results

We test how well existing language models can address the task of logical fallacy classification, and check whether our proposed model can lead to performance improvement.

Zero-Shot Classifiers. In Table 5, we first look into some commonly used off-the-shelf zero-shot classification models. Surprisingly, most zero-shot classifiers are not much better than randomly choosing a label (i.e., the "Random" baseline in Table 5). The RoBERTa-MNLI classifier and GPT2, which achieve 12.22% and 13.67% F_1 scores, respectively are only marginally better than random guessing.

Finetuned Models. We further look into the effectiveness of finetuned large language models. The model performance is shown in ascending order in Table 5. According to our main metric, F_1, the best language model is Electra, which achieves 53.31% F_1 scores, followed by DeBERTa which achieves 50.29%.

Then, we adopt Electra as the backbone model to test our proposed structure-aware classifier (denoted as Electra-*StructAware*). Our model outperforms Electra by 5.46%, which is a fairly large margin. This implies the importance of encouraging the model to shift its attention to the logical form. Our model also achieves the highest exact match result, 47.67%, which is 12.01% better than the best performance among all language models finetuned in the standard way.

	F_1	P	R	Freq.
Faulty Generalization	60.24	47.62	81.97	18.01
Ad Hominem	78.65	72.92	85.37	12.33
Ad Populum	79.45	67.44	96.67	9.47
False Causality	58.82	62.50	55.56	8.82
Circular Claim	46.43	35.14	68.42	6.98
Appeal to Emotion	50.00	48.00	52.17	6.82
Fallacy of Relevance	39.22	37.04	41.67	6.61
Deductive Fallacy	25.81	16.67	57.14	6.21
Intentional Fallacy	26.23	17.39	53.33	5.84
Fallacy of Extension	49.18	37.50	71.43	5.76
False Dilemma	55.00	39.29	91.67	5.76
Fallacy of Credibility	58.82	58.82	58.82	5.39
Equivocation	33.33	100.00	20.00	2.00
Overall	58.77	55.25	63.67	100

Table 6: Class-specific performance achieved by Electra-*StructAware*. For each class, we report the F_1 score, precision (P), recall (R), and the frequency (Freq.) of the class in the LOGIC dataset. Note that the Freq. column is copied from Table 1.

Ablation Study. Through the ablation study in Table 5, we can see that raw premise (i.e., keeping the original text input) with structure-aware hypothesis yields 56.72%, which can be attributed to the fact that the logical form provides richer information than just the label name. On the contrary, the structure-aware premise with the raw hypothesis of just the label name leads to a much worse result, perhaps because the model cannot easily figure out the correspondence between the masked text input and the label name. The ablation study also demonstrates that the best performance of our model comes from the matching between the logical form and the masked text input.

4.3 Class-Specific Performance

In addition to the overall performance of our proposed Electra-*StructAware* model, we further analyze its class-specific performance in Table 6.

Many of the logical fallacy classes can reach F_1 scores close to the overall F_1 of 58.77%. However, there are some logical fallacy types with relatively higher or lower performance. As the prediction performance can depend on both the difficulty of

	F_1	P	R
Direct Transfer			
Electra	22.72	18.68	35.85
Electra-*StructAware*	**27.23**	20.46	45.12
Finetuned further on **LOGICCLIMATE**			
Electra (Ft)	23.71	20.86	23.09
Electra-*StructAware* (Ft)	**29.37**	17.66	67.22

Table 7: Performance of *direct transfer* models trained on LOGIC and tested on LOGICCLIMATE. We also include additional results of the same two models further finetuned and tested on LOGICCLIMATE. Since LOGICCLIMATE is a multi-label classification, we omit the accuracy as it is not applicable here.

identifying a logical fallacy type as well as the number of training samples for that type, we also provide the frequency (%) of each logical fallacy in Table 6.

We can notice that the best-performing classes are *ad populum* (F_1=79.45%) and *ad hominem* (F_1=78.65%), which even outperform the most frequent class, *faulty generalization* (F_1=60.24%). A possible reason can be that *ad populum* can be detected often when there are numbers or terms that refer to a majority of people, and *ad hominem* uses insulting words or undermines the credibility of a person.

We further look into logical fallacies that are difficult to learn. For example, among the four logical fallacies with a similar frequency of 6+% in the dataset, namely *circular claim*, *appeal to emotion*, *fallacy of relevance*, and *deductive fallacy*, the one that is the most difficult to learn is *deductive fallacy* (F_1=25.81%), which has the lowest F_1 across all 13 classes. This might be a combined effect of the difficulty of distilling the formal logic from various content words in this case, and also that there can be several more forms of deductive fallacies which are not covered by our approach. This could be an interesting direction for future work.

4.4 Extrapolating to LOGICCLIMATE

We also test our models on the more challenging test set, LOGICCLIMATE, to check how well the models can extrapolate to an unseen domain, namely claims in climate change news articles. We use the two best-performing models trained on LOGIC, namely the best language model Electra and our proposed Electra-*StructAware* model.

In Table 7, the direct transfer performance is calculated by directly using the two models trained on LOGIC and testing them on the entire LOGICCLIMATE. Although both models drop drastically

Correct Predictions
"You should drive on the right side of the road because that is what the law says, and the law is the law." **Ground-truth label:** Circular claim
"Some kangaroos are tall. Some MMA fighters are tall. Therefore, some kangaroos are MMA fighters." **Ground-truth label:** Deductive fallacy
Incorrect but Reasonable Predictions
"Drivers in Richmond are terrible. Why does everyone in a big city drive like that?" **Ground-truth label:** Ad hominem **Predicted label:** Faulty generalization
"Whatever happens by chance should be punished because departure from laws should be punished." **Ground-truth label:** Equivocation **Predicted label:** Circular claim
Incorrect Predictions
"**A** car makes less pollution than **a** bus. Therefore, car**s** are less of a pollution problem than bus**es**." **Ground-truth label:** Faulty generalization **Predicted label:** Circular claim
"Not that it ever was a thing, really. This debate – as I argue at some length in Watermelons – was always about **left-wing ideology, quasi-religious hysteria, and 'follow the money' corruption**, never about 'science.' Still, it's always a comfort to know that 'the science' is on our side too. They do so hate that fact, **the Greenies**." **Ground-truth label:** Ad hominem and the fallacy of extension **Predicted label:** Intentional fallacy

Table 8: Examples of correct predictions, incorrect but reasonable predictions, and incorrect predictions.

when transferring to the unseen LOGICCLIMATE challenge set, our model Electra-*StructAware* achieves the higher performance, 27.23%, and still keeps its relative improvement of 4.51% over the Electra baseline.

We also include an additional experiment of finetuning the two models on LOGICCLIMATE, where both show improvements, and Electra-*StructAware* outperforms Electra by a larger margin of 5.66%. The detailed setup of this additional experiment is in Appendix C.2. As we can see, even the finetuned numbers are still lower than those of LOGIC, so we encourage more future work to enhance the out-of-domain generalizability of logical fallacy classifiers.

4.5 Error Analysis

Next, we analyze our model predictions and common error types. We identify three categories of model predictions in Table 8: correct predictions, incorrect but reasonable predictions, and incorrect predictions. Common among incorrect but reasonable predictions are some debatable cases where multiple logical fallacy types seem to apply, and the ground-truth label marks the most obvious one.

For example, "Drivers in Richmond are terrible. Why does everyone in a big city drive like that?" is an example of ad hominem as it is a personal attack against drivers in Richmond, but also has some flavor of faulty generalization from "drivers in Richmond" to "everyone in a big city."

Among the incorrect predictions, we can see the difficulty of identifying the nuances in the logical forms. The sample from LOGIC, "A car makes less pollution than a bus. Therefore, cars are less of a pollution problem than buses.", at first glance, looks similar to circular reasoning as it seems to repeat the same argument twice. However, in fact, it is a faulty generalization from "a car. . . a bus" to "cars. . . buses." Another sample from LOGICCLIMATE uses context-specific words "left-wing ideology, quasi-religious hysteria, and 'follow the money' corruption. . . the Greenies" for ad hominem when politically criticizing climate change advocates.

5 Limitations and Future Work

Some limitations of the current proposed model is that it can be effective for text with clear spans of paraphrases, but does not always work for more complicated natural text, such as the journalistic style in the climate change news articles. Another limitation is that, in the scope of this work, we only explored one logical form for each fallacy type. Since there could be multiple ways to verbalize each fallacy, future work can explore if the models can match the input text to several candidate logical forms, and create a multi-way voting system to decide the most suitable logical fallacy type.

Orthogonal to model development, future work can also explore other socially meaningful applications of this work, in line with the NLP for Social Good Initiative (Jin et al., 2021; Gonzalez et al., 2022),[3] logical fallacy detection can be used in various settings: to validate information and help fight misinformation along with fact-checkers (Riedel et al., 2017; Thorne et al., 2018), to check whether cognitive distortions (Beck, 1963; Kaplan et al., 2017; Lee et al., 2021) are correlated with some types of logical fallacies, to check whether some logical fallacies are commonly used as political devices of persuasion in politicians' social media accounts, among many other possible application cases.

[3] https://nlp4sg.vercel.app

6 Related Work

Logical Fallacies. Logic in language is a subject that has been studied since the time of Aristotle, who considers logical fallacies as "deceptions in disguise" in language (Aristotle, 1991). Logical fallacies refer to errors in reasoning (Tindale, 2007), and they usually happen when the premises are not relevant or sufficient to draw the conclusions (Johnson and Blair, 2006; Damer, 2009). Early studies on logical fallacies include the taxonomy (Greenwell et al., 2006), general structure of logical arguments (Toulmin, 2003), and schemes of fallacies (Walton et al., 2008).

Logic is at the center of research on argumentation theory, an active research field in both the linguistics community (Damer, 2009; Van Eemeren et al., 2013; Govier, 2013), and NLP community (Wachsmuth et al., 2017b,a; Habernal et al., 2018a; Habernal and Gurevych, 2016). The most relevant NLP works include classification of argument sufficiency (Stab and Gurevych, 2017), ad hominem fallacies from Reddit posts (Habernal et al., 2018b) and dialogs (Sheng et al., 2021), as well as automatic detection of logical fallacies using a rule parser (Nakpih and Santini, 2020).

To the best of our knowledge, our work is the first to formulate logical fallacy classification with deep learning models, and also the first to propose logical fallacy detection for climate change news.

Combating Misinformation. There has been an increasing trend of using NLP to combat misinformation and disinformation (Feldman et al., 2019). Most existing works focus on fact-checking, which uses evidence to verify a claim (Pérez-Rosas et al., 2018; Thorne et al., 2018; Riedel et al., 2017). To alleviate the computationally expensive fact-checking procedures against external knowledge sources, some other efforts include check-worthy claim detection (Konstantinovskiy et al., 2018), out-of-context misinformation detection (Aneja et al., 2021), while some still need to outsource to manual efforts (Nakov et al., 2021). We consider our work of logical fallacy detection to be independent of the topic and content, which can be an orthogonal component to existing fact-checking work. The logical fallacy checker can be used before or along with fact-checkers to reduce the number of claims to check against, by eliminating logically fallacious claims in the first place. Logical fallacies also have some intersections with propaganda techniques (Da San Martino et al., 2019b,a, 2020a,b), but they are two distinct tasks, since propaganda is more about influencing people's mindsets and the means can be various types of persuasion devices, and this work on logical fallacies mainly focuses on the logical and reasoning aspect of language, with implications for enhancing the reasoning ability of NLP models.

7 Conclusion

This work proposed logical fallacy detection as a novel task, and constructed a dataset of common logical fallacies and a challenge set of fallacious climate claims. Using this dataset, we tested the performance of 12 existing pretrained language models, which all have limited performance when identifying logical fallacies. We further proposed a structure-aware classifier which surpasses the best language model on the dataset and the challenge set. This dataset provides a ground for future work to explore the reasoning ability of NLP models.

Acknowledgments

We thank Kevin Jin for insightfully pinpointing the prevalence of logical fallacies in discussions of social problems. We thank the labmates at the LIT Lab at the University of Michigan, especially Ashkan Kazemi for constructive suggestions and writing advice based on existing work in fake news detection. We thank Prof Markus Leippold (University of Zürich) for insights on climate change fact verification datasets. We thank Amelia Francesca Hardy (Stanford) for discussions on pressing social problems that NLP can be promising to address.

We especially thank many annotators at the University of Michigan for helping us with the dataset, including Safiyyah Ahmed, Jad Beydoun, Elizabeth Loeher, and Brighton Pauli. Additional thanks to Jad Beydoun for helping to compile some numbers and examples in this paper.

This material is based in part upon works supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by the Precision Health Initiative at the University of Michigan; by the John Templeton Foundation (grant #61156); and by a Responsible AI grant by the Haslerstiftung.

Ethical Considerations

The data used in this work are all from public resources, with no user privacy concerns. The potential use of this work is for combating misinformation and helping to verify climate change claims.

Contributions of the Authors

This project was a large collaboration that would not have happened without dedicated effort from every co-author.

The *idea of the project* originated in discussions among Zhijing Jin, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan.

For the *dataset collection*, Zhijing Jin led the data collection. She conducted the annotation and compilation, together with Yvonne Ding who collected the original articles for LOGICCLIMATE, as well as Zhiheng Lyu who automatic crawled part of LOGIC.

Analyses of dataset characteristics and experimental results were first done by Zhijing Jin, and later updated by Abhinav Lalwani. Some analyses in the appendix were done by Zhiheng Lyu.

For the *experiments*, the first round was done by Tejas Vaidhya, the second round was done by Zhijing Jin and Xiaoyu Shen, and the final round was done by Abhinav Lalwani, including the Electra-*StructAware*.

Cleaning and compilation of the code and data was done by Zhijing Jin and then updated by Abhinav Lalwani.

All co-authors contributed to *writing the paper*, especially Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, Xiaoyu Shen, and Bernhard Schoelkopf.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: an easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 54–59. Association for Computational Linguistics.

Shivangi Aneja, Christoph Bregler, and Matthias Nießner. 2021. Catching out-of-context misinformation with self-supervised learning. *CoRR*, abs/2101.06278.

Aristotle. 1991. *On Rhetoric: A Theory of Civil Discourse*. Oxford University Press.

Aristotle. 2006. *On sophistical refutations*. The Internet Classics Archive.

Aaron T Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333.

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

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. 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.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020b. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. 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.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Reasoning*. Wadsworth Cengage Learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Anna Feldman, Giovanni Da San Martino, Alberto Barrón-Cedeño, Chris Brew, Chris Leberknight, and Preslav Nakov, editors. 2019. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. Association for Computational Linguistics, Hong Kong, China.

Dov M Gabbay and John Hayden Woods. 2004. *Handbook of the History of Logic*, volume 2009. Elsevier North-Holland.

Fernando Gonzalez, Zhijing Jin, Jad Beydoun, Bernhard SchÃ¶lkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2022. How is NLP addressing the UN Sustainable Development Goals? a challenge set to analyze NLP for social good papers.

Trudy Govier. 2013. *A practical study of argument*. Cengage Learning.

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 (ISSC)*.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. 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)*, pages 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. 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.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020a. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3202–3213. International Committee on Computational Linguistics.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020b. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? A sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, pages 3099–3113. Association for Computational Linguistics.

Ralph Henry Johnson and J Anthony Blair. 2006. *Logical self-defense*. International Debate Education Association.

Simona C Kaplan, Amanda S Morrison, Philippe R Goldin, Thomas M Olino, Richard G Heimberg, and James J Gross. 2017. The cognitive distortions questionnaire (cd-quest): validation in a sample of adults with social anxiety disorder. *Cognitive therapy and research*, 41(4):576–587.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *CoRR*, abs/1809.08193.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for efficient, explainable, and reusable systems: A case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4257–4272, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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. *CoRR*, abs/1907.11692.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.

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

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *CoRR*, abs/1707.03264.

Bertrand Russell. 2013. *History of western philosophy: Collectors edition*. Routledge.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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.

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.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: A compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2158–2170. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Frans H Van Eemeren, Rob Grootendorst, Ralph H Johnson, Christian Plantin, and Charles A Willard. 2013. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments*. Routledge, Taylor & Francis Group.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

A More Details of the Dataset

A.1 Dataset Overview for Responsible NLP

Documentation of the artifacts:.

- Coverage of domains: general domain (e.g., educational examples of logical fallacies), and climate change news articles with logical fallacies.
- Languages: English.
- Linguistic phenomena: Logical fallacies.
- Demographic groups represented: No specific demographic groups.

Annotation details:.

- Basic demographic and geographic characteristics of the annotator population that is the source of the data: All annotators are native English speakers who are undergraduates at a university in the US. There are two male annotators and two female annotators.
- How you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence): We broadcast the recruitment to the undergraduate CS student mailing list at a university. We received a large number of applications and selected four annotators. We followed the university's standard payment of 14 USD/hour for each student.
- How consent was obtained from annotators: We explained to the annotators that the data will be open-sourced for research purpose.
- Data collection protocol approved (or determined exempt) by an ethics review board: The dataset included in this work did not go through reviews by an ethics review board.
- Full text of instructions given to participants: We first show to the participants the description and examples of the 13 logical fallacy types as in Appendix D, and when they are actually annotating, the interface screenshots are in Figures 3 and 4.

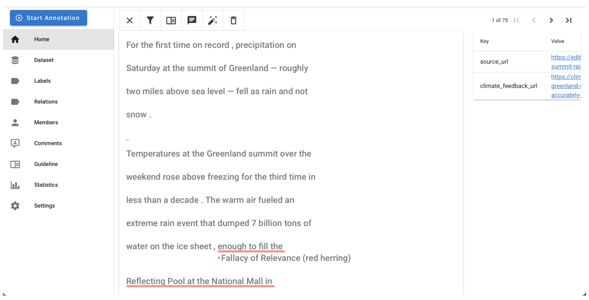

Figure 3: Annotation interface for the LogicClimate challenge set.

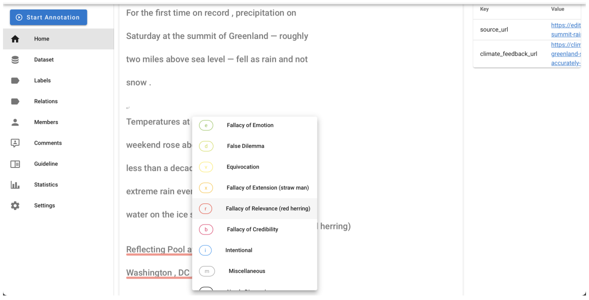

Figure 4: Choices of logical fallacy types in the annotation interface of the LogicClimate challenge set.

Data sheet:.

- Why was the dataset created: We created the dataset for the proposed logical fallacy classification task.
- Who funded the creation of the dataset: The Logic part was collected by the co-authors, and the LogicClimate part was collected using the funding of a professor at the university.
- What preprocessing/cleaning was done: We tokenized the text using the word tokenization function of NLTK.[4]
- Will the dataset be updated; how often, by whom: No, the dataset will be fixed.

Additional ethical concerns:.

- Whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content: No, the dataset does not contain personal information.
- License or terms for use and/or distribution: The dataset is open-sourced with the MIT license, and the intended use is for academic research but not commercial purposes.

A.2 Data Filtering Details of Logic

The data automatically crawled from quiz websites contain lots of noises, so we conducted multiple filtering steps. The raw crawling by keyword matching such as "logic" and "fallacy" gives us 52K raw, unclean data samples, from which we filtered to 1.7K clean samples.

As not all of the automatically retrieved quizzes are in the form of "Identify the logical fallacy in this example: [...]", we remove all instances where the quiz question asks about irrelevant things such as the definition of a logical fallacies, or quiz questions with the keyword "logic" but in the context of other subjects such as logic circuits for electrical engineering, or pure math logic questions. This is

[4] https://nltk.org/

Fallacy Name	Description	Logical Form
Faulty Generalization	An informal fallacy wherein a conclusion is drawn about all or many instances of a phenomenon on the basis of one or a few instances of that phenomenon. is an example of jumping to conclusions.	[MSK1] has attribute [MSK2]. [MSK1] is a subset of [MSK3]. Therefore, all [MSK3] has attribute [MSK2]. (Reference)
False Causality	A statement that jumps to a conclusion implying a causal relationship without supporting evidence	[MSK1] occurred, then [MSK2] occurred. Therefore, [MSK1] caused [MSK2]. (Reference)
Circular Claim	A fallacy where the end of an argument comes back to the beginning without having proven itself.	[MSK1] is true because of [MSK2]. [MSK2] is true because of [MSK1]. (Reference)
Ad Populum	A fallacious argument which is based on affirming that something is real or better because the majority thinks so.	A lot of people believe [MSK1]. Therefore, [MSK1] must be true. (Reference)
Ad Hominem	An irrelevant attack towards the person or some aspect of the person who is making the argument, instead of addressing the argument or position directly.	[MSK1] is claiming [MSK2]. [MSK1] is a moron. Therefore, [MSK2] is not true. (Reference)
Deductive Fallacy	An error in the logical structure of an argument.	If [MSK1] is true, then [MSK2] is true. [MSK2] is true. Therefore, [MSK1] is true. (Reference)
Appeal to Emotion	Manipulation of the recipient's emotions in order to win an argument.	[MSK1] is made without evidence. In place of evidence, emotion is used to convince the interlocutor that [MSK1] is true. (Reference)
False Dilemma	A claim presenting only two options or sides when there are many options or sides.	Either [MSK1] or [MSK2] is true. (Reference)
Equivocation	An argument which uses a key term or phrase in an ambiguous way, with one meaning in one portion of the argument and then another meaning in another portion of the argument.	[MSK1] is used to mean [MSK2] in the premise. [MSK1] is used to mean [MSK3] in the conclusion. (Reference)
Fallacy of Extension	An arugment that attacks an exaggerated or caricatured version of your opponent's position.	[MSK1] makes claim [MSK2]. [MSK3] restates [MSK2] (in a distorted way). [MSK3] attacks the distorted version of [MSK2]. Therefore, [MSK2] is false. (Reference)
Fallacy of Relevance	Also known as red herring, this fallacy occurs when the speaker attempts to divert attention from the primary argument by offering a point that does not suffice as counterpoint/supporting evidence (even if it is true).	It is claimed that [MSK1] implies [MSK2], whereas [MSK1] is unrelated to [MSK2] (Reference)
Fallacy of Credibility	An appeal is made to some form of ethics, authority, or credibility.	[MSK1] claims that [MSK2]. [MSK1] are experts in the field concerning [MSK2]. Therefore, [MSK2] should be believed. (Reference)
Intentional Fallacy	A custom category for when an argument has some element that shows intent of a speaker to win an argument without actual supporting evidence.	[MSK1] knows [MSK2] is incorrect. [MSK1] still claim that [MSK2] is correct using an incorrect argument.

Table 9: Types of logical fallacies along with their descriptions and logical forms.

done by writing several matching patterns. After several processing steps such as deleting duplicates, we end up with 7,389 quiz questions. Moreover, as there is some noise that cannot be easily filtered by pattern matching, we also manually go through the entire dataset to only keep sentences that contain examples of logical fallacies, but not other types of quizzes.

The entire cleaning process resulted in 1.7K high-quality logically fallacious claims in our dataset. As a reference, for each fallacy example we also release the URL of the source website where we extract this example from.

A.3 Logical Fallacy Types

As different sources use different names for logical fallacies, we composed a set of 13 logical fallacy categories by conforming to set of logical fallacies given by Wikipedia,[5] and considering the most common types in the dataset. Therefore, we merged different surface forms of the same logical fallacy by listing out the different names of the same logical fallacy introduced on Wikipedia and also provided by educational websites. This leads to a reduction in logical fallacy types. For example, "hasty induction" and "jumping to conclusions" are merged under the category of "hasty generalization." We further improve the eventual list by handcrafted rules, and delete data samples that cannot be matched to any of the logical fallacy types in our list. For a small number of remaining logical fallacy names which we cannot resolve au-

[5] https://en.wikipedia.org/wiki/List_of_fallacies

tomatically, we ask human annotators to align the names.

We introduce the detailed description for each of the 13 logical fallacy types and their logical forms in Table 9. Most the descriptions and logical forms are collected from online websites introducing these logical fallacies. We provide the link of the source websites as references, and in some cases, we paraphrase the logical form to make it closer to natural text. In addition, we also provide the introduction of the 13 types that we compiled for the annotators in Appendix D.

A.4 Data Annotation Details of LogicClimate

To ensure good annotation quality, we ask the annotators to pass a test batch of 65 samples after reading the definitions and examples of the 13 logical fallacies in the Logic dataset carefully. The test batch consists of 5 randomly selected samples for each of the 13 logical fallacies, and the annotators achieve above 85% accuracy. We explained the examples where they did incorrectly and resolved their questions before they started annotating the LogicClimate test set.

Since each sample is annotated by two different annotators, we finalize the ground-truth labels in the following way: The two annotators merge all their annotations, and for places with divergent opinions, they cross-check with the experts' written reviews on the Climate Feedback website for each article. Specifically, each article is commented on by multiple expert reviewers such as professors, senior scientists and other researchers who explain what is fallacious with the article. If the labels can still not be unified after checking the expert reviews, since the annotators are trained to master the tasks very well (with 5+ hours of training, testing and discussions before the annotation), we let the two annotators have a discussion to decide the final label.

A.5 LogicClimate Examples

We also show examples of LogicClimate in Table 10.

B Implementation Details

B.1 Details of Zero-Shot Baselines

For the zero-shot classification models, we used pretrained NLI models (Yin et al., 2019) that are default choices of zero-shot classifiers in the transformers library (Wolf et al., 2020): BART-MNLI and RoBERTa-MNLI. For the implementation, we follow the standard pipeline introduced by huggingface.[6]

For the TARS model (Halder et al., 2020b), we follow the official documentation.[7] All the above zero-shot classifiers show reasonable performance on existing datasets.[8]

For GPT-3, we follow the official guide,[9] and use the following prompt (without additional efforts on prompt tuning because we do not assume any training samples for the zero-shot classification model): "Please classify a piece of text into the following categories of logical fallacies: [a list of all logical fallacy types].

Text: [Input text]

Label:"

We use the default search engine "davinci," and the model "curie." For reproducibility, we set the temperature to 0 for GPT-3 and all our zero-shot classification codes use a random seed of 1.

B.2 Details of Finetuned Models

All models are finetuned using the NLI task, as motivated in Section 3.1. We used a learning rate of 2^{-5}, and the AdamW optimizer. We do not turn the learning rate and optimizers since the loss converges smoothly in all cases. We tune the weights of the positive class, and set it to be 12, and the negative classes to be 1. The models used in our experiments have between 11M and 140M parameters. We train all models using NVIDIA TITAN RTX machines for less than two GPU Hours. For reproducibility, we fix the random seed to zero, and report the statistics of a single run.

Due to the different dataset nature, our Logic is a single-label multi-class classification and the LogicClimate is a multi-label multi-class classification.

B.3 Details of Our Structure-Aware Model

For the structure-aware classifier, we set the threshold of cosine similarity between two text spans to be 0.7, which is tuned using a manual grid search

[6] https://bit.ly/3E92Mvq

[7] https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_10_TRAINING_ZERO_SHOT_MODEL.md

[8] https://github.com/nlptown/nlp-notebooks/blob/master/Zero-Shot%20Text%20Classification.ipynb

[9] https://beta.openai.com/docs/api-reference/classifications

Logical Fallacy	Examples
Faulty Generalization	"**For decades** horticulturalists have pumped carbon dioxide into glasshouses to increase yields. The fossil **record shows** that a thriving and diversification of plant and animal life **occurs every time** the atmosphere had a very high carbon dioxide content. **In the past, warming has never been a threat** to life on Earth."
Ad Hominem	"While CO2 levels were continuing to rise, temperatures weren't. Hence the need for a **fallback position** — an environmental theory which would justify the massively expensive and disruptive ongoing decarbonisation programme so **assiduously championed** by politicians, scientists, green campaigners and anyone **making money out of the renewables business**. Ocean acidification fitted the bill perfectly."
Ad Populum	"According to a recent National Economic Research Associates Economic Consulting study, the Paris Agreement **could obliterate $3 trillion** of GDP, **6.5 million industrial sector jobs** and **$7,000 in per capita household income** from the American economy by 2040. Meeting the 2025 emissions reduction target alone **could subtract $250 billion from our GDP** and **eliminate 2.7 million jobs**. The cement, iron and steel, and petroleum refining industries could see their **production cut by 21% 19%, and 11%** respectively."
False Causality	"But like most claims regarding global warming, the real effect is small, **probably** temporary, and **most likely due** to natural weather patterns. Any changes in hurricanes over 70 years, even if real, **can easily be part of** natural cycles — or incomplete data. Coastal lake sediments along the Gulf of Mexico shoreline from 1,000 to 2,000 years ago **suggest** more frequent and intense hurricanes than occur today."
Circular Claim	"Even **if** enough accurate surface temperature measurements existed to ensure reasonable planetary coverage (**it doesn't**) and to calculate some sort of global temperature statistic, interpreting **its significance would be challenging**. What averaging rule would you use to handle the data from thousands of temperature-sensing stations?"
Appeal to Emotion	"There are now, trapped in Arctic ice, **diseases** that have not circulated in the air for millions of years — in some cases, since before humans were around to encounter them. Which means our immune systems would have **no idea how to fight back** when those prehistoric **plagues emerge** from the ice."
Fallacy of Relevance	"But **there are also reasons to believe** that environmental alarmism will, if not come to an end, have diminishing cultural power. The coronavirus pandemic **is an actual** crisis **that puts** the climate "crisis" **into perspective**. Even if you think we have overreacted, Covid-19 has killed nearly 500,000 people and shattered economies around the globe."
Deductive Fallacy	"Indeed, Queensland's 2014 heat wave paled in comparison to the 1972 heat wave that occurred 42 years of global warming ago. **If** global warming **caused** the 2014 Queensland heat wave, **why wasn't** it as severe as the 1972 Queensland heat wave? **Blaming every** single summer heat wave or extreme weather **event** on global warming is a stale and discredited tactic in the alarmist playbook."
Intentional Fallacy	"The bottom line is **there's no solid connection** between climate change and many major indicators of extreme weather that politicians keep talking about, such as hurricanes, tornadoes, droughts, rainfall and floods, despite Trudeau's claims to the contrary. The continual **claim of such links is misinformation** employed for political and rhetorical purposes."
Fallacy of Extension	"**For global warming alarmists**, however, a greener biosphere **is terrible** news and something **to be opposed**. This, in a nutshell, **defines the opposing sides** in the global warming debate. **Global warming alarmists claim** a greener biosphere with richer and more abundant plant life is horrible and justifies massive, economy-destroying energy restrictions. Global warming realists understand that a greener biosphere with richer and more abundant plant life **is not a horrible thing** simply because humans may have had some role in creating it."
False Dilemma	"America is poised to become a net energy exporter over the next decade. We **should not abandon that** progress **at the cost of** weakening our energy renaissance and crippling economic growth."
Fallacy of Credibility	"**I note particularly** that sea-level rise is not affected by the warming; it continues at the same rate, 1.8 millimeters a year, **according to** a 1990 review by Andrew S. Trupin and John Wahr. **I therefore conclude**—contrary to the general wisdom—that the temperature of sea water has no direct effect on sea-level rise. That means neither does the atmospheric content of carbon dioxide."
Equivocation	"Also, the alarmist assertion that polar ice sheets are melting is simply false. Although alarmists frequently point to a modest recent **shrinkage in the Arctic** ice sheet, that decline has been completely offset by ice sheet **expansion in the Antarctic**. **Cumulatively, polar ice sheets have not declined** at all since NASA satellite instruments began precisely measuring them 35 years ago."

Table 10: Examples of the LOGICCLIMATE data.

based on performance on the dev set. In the training, we keep the training samples of the original text, and add additional samples using the masked text; in the inference stage, we choose the input format that performs the better on the development set, which is the masked text format.

C Additional Experiments

C.1 Class-Specific Performance on LOGICCLIMATE

For LOGICCLIMATE, we provide class-specific performance for the best performing model in Table 11. There is lots of space for future work to improve the performance on this dataset. For error analysis with the current best performing model, we identify the following aspects that make LOGICCLIMATE more challenging than LOGIC. We measure the complexity and diversity of the dataset by measuring the BLEU score difference with the logical forms, and find that LOGICCLIMATE (0.18) has a lower similarity than LOGIC (0.24). The relatively higher complexity and diversity might explain why the best model only achieves 29.37% on LOGICCLIMATE. We also find that as LOGIC has examples that are designed such that students can classify them, LOGICCLIMATE has fallacies created by top-level journalists created with the intention that even educated readers will not be able to detect them. The small size of the dataset might also be a factor, as we find that the best performing model achieves a similar performance (34.52%) on LOGIC when trained on the same amount of data. We also find that the model struggles on classes with small amounts of data.

C.2 Finetuning on LOGICCLIMATE

To obtain the performance of Electra-*StructAware* vs. Electra after finetuning on the LOGICCLIMATE dataset, we split the LOGICCLIMATE dataset into train, dev, and test splits. Dataset statistics are shown in Table 12.

D Details of All Fallacy Types

We list the details of all fallacy types below. We also use this list to guide annotators to identify

	F_1	P	R	Freq.
Intentional	24.58	100.00	39.46	25.58
Appeal to Emotion	23.40	84.62	36.67	11.37
Faulty Generalization	16.56	96.43	28.27	10.18
Fallacy of Credibility	25.00	45.00	32.14	9.90
Ad Hominem	41.67	66.67	51.28	7.84
Fallacy of Relevance	12.73	31.82	18.18	7.80
Deductive Fallacy	9.32	64.71	16.30	6.50
False Causality	15.15	31.25	20.41	5.11
Fallacy of Extension	0.00	0.00	0.00	4.91
Ad Populum	0.00	0.00	0.00	4.55
False Dilemma	16.67	16.67	16.67	3.80
Equivocation	5.00	20.00	8.00	1.94
Circular Claim	0.00	0.00	0.00	0.51
Overall	**29.37**	17.66	67.22	8.37

Table 11: Class-specific performance achieved by Electra-*StructAware* on LOGICCLIMATE. For each class, we report the F_1 score, precision (P), recall (R), and the frequency (Freq.) of the class in the LOGICCLIMATE dataset. Note that the Freq. column is copied from Table 4.

	# Samples	# Sents	# Tokens	Vocab
Total Data	1,079	1,463	38,828	5,809
Train	680	891	24,814	4,402
Dev	219	331	8,419	2,229
Test	180	241	5,595	1,707

Table 12: Statistics of the LOGIC dataset.

logical fallacies.

Faulty Generalization

- **Definition:** This fallacy occurs when an argument applies a belief to a large population without having a large enough sample to do so.
- **Example:** A New York driver cuts you off in traffic. You then decide that all New Yorkers are terrible drivers.
- **Synonyms or Subtypes:** Slippery Slope, Hasty Generalization, Accident, Fallacy of Division, Error of Division, Error of Composition, Property in the Whole, Property in the Parts, Causal Oversimplification, Part to Whole, Association Fallacy, Guilt by Association, Composition Fallacy, Ecological Fallacy, Conjunction Fallacy, False Analogy, Inconsistent Comparison, Package Deal, Overwhelming Exception, False Equivalence, All Things Are Equal, McNamara Fallacy.

False Causality

- **Definition:** This fallacy occurs when an argument assumes that since two events are correlated, they must also have a cause and effect relationship.
- **Example:** We observed an increase in ice cream sales at the same time as air conditioner sales increased. Therefore, we can conclude that selling more ice cream causes more air conditioners to be sold.
- **Synonyms or Subtypes:** Post hoc ergo propter hoc, Cum hoc ergo propter hoc, Regression Fallacy, Consecutive Relation, Magical Thinking, Gambler's Fallacy (rarely called temporal flaw/temporal fallacy), Ludic Fallacy.

Circular Claim

- **Definition:** This fallacy occurs when an argument uses the claim it is trying to prove as proof that the claim is true.
- **Example:** You must obey the law, because it is illegal to break the law.
- **Synonyms or Subtypes:** Circular Reasoning, Homunculus Fallacy.

Ad Populum

- **Definition:** This fallacy occurs when an argument is based on affirming that something is true because a statistical majority believes so.
- **Example:** Most people believe that there is a God, therefore it must be true.
- **Synonyms or Subtypes:** Appeal to the Public, Ad Numerum, Appeal to the Numbers, Bandwagon Fallacy.

Ad Hominem

- **Definition:** This fallacy occurs when a speaker trying to argue the opposing view on a topic makes claims against the other speaker instead of the position they are maintaining.
- **Example:** Person A makes a claim. Person B says that Person A's claim is false because Person A is not a hard worker.
- **Synonyms or Subtypes:** Genetic Fallacy, Tu quoque (you too), Bulverism, Poisoning the Well, Appeal to Hypocrisy, Traitorous Critic.

Deductive Fallacy

- **Definition:** This fallacy occurs when there is a logical flaw in the reasoning behind the argument, such as a propositional logic flaw.

- **Example:**
  - Affirming the consequent: If A is true then B is true. B is true. Therefore, A is true.
  - Denying the antecedent: If A is true then B is true. A is false. Therefore, B is false.
  - Affirming a disjunct: A or B is true. B is true. Therefore, A is not true.
- **Synonyms or Subtypes:** False Analogy, Affirming the Consequent, Non-sequitur, Four Terms Fallacy, Affirming the Disjunct, Argument From Fallacy (correct identification of fallacy, but incorrect conclusion), Appeal to Probability, Undistributed Middle, Moral Equivalence, Self contradiction, Internal Contradiction, Masked-man Fallacy, Four Terms, Illicit Major, Illicit Minor, Denying the Antecedent, Existential Fallacy, Kettle Logic, Affirmative Conclusion from a Negative Premise, Negative Conclusion from a Negative Premise, Exclusive Premises.

Appeal to Emotion

- **Definition:** This fallacy is when emotion is used to support an argument, such as pity, fear, anger, etc.
- **Example:** You should marry me. I know we're not compatible, but you're my last chance.
- **Synonyms or Subtypes:** Appeal to Pity, Appeal to Fear, Ad baculum (appeal to force), Appeal to Ridicule, Appeal to Gallery, Wishful Thinking, Appeal to Consequences, Appeal to Spite, Appeal to Force, Appeal to Flattery.

False Dilemma

- **Definition:** This fallacy is when incorrect limitations are made on the possible options in a scenario when there could be other options.
- **Example:** You're either for the war or against the troops.
- **Synonyms or Subtypes:** Either/Or thinking, Black-or-White Fallacy, False Dichotomy, Nirvana Fallacy, Perfect Solution.

Equivocation

- **Definition 1:** This fallacy can occur in two ways: the first is when ambiguous or evasive language is used to avoid committing oneself to a position.
- **Example:**

Speaker 1: Did you torture the prisoner?

Speaker 2: No, we just held him under water for a while, and then did a mock hanging.

- **Definition 2:** The second way equivocation occurs is when the same is word is used in an argument but with different meanings:
- **Example 2:**

Speaker 1: We are using thousands of people to go door to door and help spread the word about social injustice and the need for change.

Speaker 2: I can't be a part of this because I was taught that using people is wrong.

- **Definition 3:** An equivocation seeks to draw comparisons between different, often unrelated things.
- **Synonyms or Subtypes:** Uncertain use of term or concept, Reification, Continuum fallacy, False attribution, Moral equivalence, Etymological Fallacy.

Fallacy of Extension

- **Definition:** Also known as straw man, this is when an argument appears to be refuted by being replaced with an argument with a similar but weaker argument.
- **Example:**

Speaker 1: I think we should have single payer, universal, healthcare.

Speaker 2: Communist countries tried that. We don't want America to be a communist country so we shouldn't have single payer healthcare.

- **Synonyms or Subtypes:** Straw man, Suppressed Correlative.

Fallacy of Relevance

- **Definition:** Also known as red herring, this fallacy occurs when the speaker attempts to divert attention from the primary argument by offering a point that does not suffice as counterpoint/supporting evidence (even if it is true).
- **Example:** We should move our office to California to expand our potential customers. And the weather is warmer there, which is all the more reason to move there.
- **Synonyms or Subtypes:** Red herring, Two wrongs make a right, Argument to moderation, Moralistic fallacy, Moral equivalence, Logic

chopping, Proof by assertion, Argument from silence, Irrelevant material, Relative privation.

Fallacy of Credibility

- **Definition:** This fallacy is when an appeal is made to some form of ethics, authority, or credibility.
- **Example:** If mailing a hand-written letter was good enough in the past, then you don't need those pesky computers (appeal to tradition).
- **Synonyms or Subtypes:** Appeal to authority, Appeal to to nature, Naturalistic fallacy, Appeal to tradition, Chronological snobbery (reverse of tradition), Appeal to novelty, Ipse dixit, Etymological fallacy, Appeal to poverty, Appeal to accomplishment.

Intentional Fallacy

- **Definition:** This is sort of a "custom-made" category for when an argument has some element that shows "intent" of a speaker to win an argument without actual supporting evidence.
- **Example:** Can you meet to discuss this tomorrow, or are you too busy slacking off? (loaded question - the person who answers with yes/no is cornered into discussing or slacking off)
- **Extra example:** A dating app matches Joe and Jane because they both love the same shows, music, and going to the beach. It did not take into account their 40 year age difference, or that Joe works overnight shifts and Jane works 9-5 (texas sharpshooter).
- **Synonyms or Subtypes:** Texas sharpshooter, Cherry picking, Mcnamara fallacy, No true scotsman, Appeal to ignorance/argument from ignorance, Complex question, Moving the goalposts, Loaded question, Special pleading, Hiding information/half truth, Many questions, Incredulity, Divine Fallacy, Quoting out of context, Shifted burden of proof, Ambiguous words or phrases.