Multitask Instruction-based Prompting for Fallacy Recognition

Tariq Alhindi[1] **Tuhin Chakrabarty**[1] **Elena Musi**[3] **Smaranda Muresan**[1,2]
[1]Department of Computer Science, Columbia University
[2]Data Science Institute, Columbia University
[3]Department of Communication and Media, University of Liverpool
{tariq, tuhin.chakr, smara}@cs.columbia.edu
elena.musi@liverpool.ac.uk

Abstract

Fallacies are used as seemingly valid arguments to support a position and persuade the audience about its validity. Recognizing fallacies is an intrinsically difficult task both for humans and machines. Moreover, a big challenge for computational models lies in the fact that fallacies are formulated differently across the datasets with differences in the input format (e.g., question-answer pair, sentence with fallacy fragment), genre (e.g., social media, dialogue, news), as well as types and number of fallacies (from 5 to 18 types per dataset). To move towards solving the fallacy recognition task, we approach these differences across datasets as multiple tasks and show how instruction-based prompting in a multitask setup based on the T5 model improves the results against approaches built for a specific dataset such as T5, BERT or GPT-3. We show the ability of this multitask prompting approach to recognize 28 unique fallacies across domains and genres and study the effect of model size and prompt choice by analyzing the per-class (i.e., fallacy type) results. Finally, we analyze the effect of annotation quality on model performance, and the feasibility of complementing this approach with external knowledge.

1 Introduction

A fallacious argument is one that seems valid but it is not (Hamblin, 2022). Theoretical work in argumentation has introduced various typologies of fallacies. For example, Van Eemeren et al. (2002) consider fallacies that occur when an argument violates the ten rules of a critical discussion, while Tindale (2007) categorizes fallacies into 4 categories: structural fallacies, related to the number and structure of arguments; fallacies from diversion, drawing from the (un)intentional diversion of the attention from the issue at hand; logical fallacies, related to the argument scheme at play and language fallacies, related to vagueness or ambiguity. Fallacious reasoning can bring misbehaviour and be used for manipulation purposes. Thus, having a system that can recognize fallacy types across domains and genres is crucial for applications that teach humans how to identify fallacies and avoid using them in their arguments.

Question-Answering dialog moves in ARGOTARIO: Has anyone been on the moon? The moon is so far away, we should focus on our society. **Fallacy:** Red Herring
Propaganda techniques in news: The ability to build an untraceable, unregistered gun is definitely a game changer. **Fallacy:** Loaded Language
Educational website on fallacies: She is the best because she is better than anyone else **Fallacy:** Circular Reasoning
Fact-checked news: Says Joe Biden has said 150 million Americans died from guns and another 120 million from COVID-19. **Fallacy:** Cherry Picking

Table 1: Examples of fallacies from multiple datasets

Work in computational models for fallacy recognition is still in its infancy, with a limited set of relatively small datasets such as ARGOTARIO (Habernal et al., 2017), which consists of question and answer dialog moves; name-calling in social media debates (Habernal et al., 2018), fallacies as propaganda techniques in news (Da San Martino et al., 2019b); logical fallacies from educational websites (Jin et al., 2022), and fallacies used for misinformation in social media and news around Covid-19 (Musi et al., 2022). Table 1, shows some examples of fallacies from these datasets.

Previous work on fallacy recognition has tackled just one dataset at a time. For example, work on detecting propaganda techniques use fine-tuning of different pre-trained transformers with embedding-based or handcrafted features (Da San Martino et al., 2020; Jurkiewicz et al., 2020) as well as

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8172–8187
December 7-11, 2022

LSTMs and transformers for sequence tagging of propaganda fragments (Da San Martino et al., 2019a; Yoosuf and Yang, 2019; Alhindi et al., 2019; Chernyavskiy et al., 2020), while Jin et al. (2022) propose a structure-aware classifier to detect logical fallacies.

Fallacy recognition is a challenging task for three main reasons: i) the number of classification labels (fallacy types) and class imbalance in existing datasets is often very high; ii) existing datasets cover varying genres and are typically very small in size due to annotation challenges; and iii) models trained on individual data sets often show poor out of distribution generalization.

A recent line of work (Wei et al., 2022; Sanh et al., 2022) relies on the intuition that most natural language processing tasks can be described via natural language instructions and models trained on these instructions in a multitask framework show strong zero-shot performance on new tasks. Based on this success, we propose a *unified model based on multitask instruction-based prompting* using T5 (Raffel et al., 2020) to solve the above challenges for fallacy recognition (Section 3). This approach allows us to unify all the existing datasets and a newly introduced dataset (Section 2) by converting 28 fallacy types across 5 different datasets into natural language instructions. In particular, we address the following research questions: i) Can we have a unified framework for fallacy recognition across domains, genres, and annotation schemes? ii) Are fallacy types expressed differently across datasets? iii) What are the effects of model size and prompt choice on the per-class performance for the fallacy recognition task?

Experimental evidence shows that our multitask fine-tuned models outperform task specific models trained on a single dataset by an average margin of 16% as well as beat strong few-shot and zero-shot baselines by average margins of 25% and 40%, respectively in macro F1 scores across five datasets (Section 4.1). To further deepen our understanding towards the task of fallacy recognition we analyze the performance of our models for each fallacy type across datasets, model size and prompt choice (Section 4.2). We further analyze the effect of annotation quality on the model performance, and the feasibility of complementing this approach with external knowledge (Section 4.3). We make all datasets, code and models publicly available.[1]

[1] https://github.com/Tariq60/fallacy-detection

Data	Ex	F	Genre	Domain
ARGOTARIO	880	5	Dialogue	General
PROPAGANDA	5.1k	15†	News	Politics
LOGIC	4.5k	13	Diverse	Education
COVID-19	621	9‡	SocMed/News	Covid-19
CLIMATE	477	9‡	News	Climate

Table 2: Summary of five fallacy datasets. **Ex**: Total number of examples. **F**: Final number of fallacy types after unifing all datasets. †Original scheme has 18 propaganda techniques. ‡Original scheme has 10 fallacy types.

2 Data

We experiment with five datasets (4 existing and a new dataset) that cover 28 unique fallacy types in multiple domains (e.g., covid-19, climate change, politics) and genres (e.g. news articles, QA turns in dialog, social media). Table 2 shows a summary of all datasets (65% training, 15% development, 20% test) and Table 8 in Appendix B shows detailed counts of fallacy types in each split for all datasets.

Existing Fallacy Datasets We include four existing fallacy datasets in our experiments.

The first dataset is ARGOTARIO, introduced by Habernal et al. (2017), a dataset for fallacy detection where given a QA pair the task is to detect the fallacy in answers. Their scheme include five fallacy types: *Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization, Irrelevant Authority*.

The second dataset (PROPAGANDA) contains 18 propaganda techniques in news articles annotated at the fragment and sentence levels (Da San Martino et al., 2019b). We focus on 15 that are fallacies and frequent enough in the data: *Loaded Language, Name Calling or Labeling, Exaggeration or Minimization, Doubt, Appeal to Fear/Prejudice, Flag-Waving, Causal Oversimplification, Slogans, Appeal to Authority, Black-and-White Fallacy, Thought-Terminating Cliche, Whataboutism, Reductio ad Hitlerum, Red Herring, and Strawman*.

The third dataset (LOGIC) is recently released by Jin et al. (2022) and contains 13 logical fallacies (*Faulty Generalization, False Causality, Circular Claim, Ad Populum, Ad Hominem, Deductive Fallacy, Appeal to Emotion, False Dilemma, Equivocation, Fallacy of Extension, Fallacy of Relevance, Fallacy of Credibility, Intentional Fallacy*) from educational websites on fallacy such as Quizziz and study.com. They contain diverse types of text such as dialogue and short statements (e.g., the *Circular*

Reasoning example shown in Table 1). The authors also introduce another challenge dataset: CLIMATELOGIC that follows the same fallacy scheme. However, it contains text segments that are too long (e.g. multiple paragraphs) with no annotations of smaller fallacious fragments like the Propaganda dataset. Therefore, CLIMATELOGIC is beyond the scope of this study.

The final existing fallacy dataset (COVID-19) is about fact-checked content around Covid-19 (Musi et al., 2022). The authors identify 10 fallacies (*Evading the Burden of Proof, Cherry Picking, Strawman, Red Herring, False Authority, Hasty Generalization, Post Hoc, False Cause, False Analogy, Vagueness*) through analysis of fact-checked social media posts and news by considering fallacies as indicators of misinformation.

More detailed description of all datasets is shown in Appendix B.

New Fallacy Dataset Drawing from the annotation scheme developed by Musi et al. (2022), we annotate 778 segments (477 fallacious) from 92 climate change articles fact-checked by climate scientists at `climatefeedback.org`. Each fact-checked article is accompanied by an "annotations" section where segments from the original articles are directly followed by the reviewers' comments. Two annotators look at both segments and comments to annotate fallacy types. They had a 0.47 Cohen's κ (Cohen, 1960), which corresponds to moderate agreement. The gold labels were then done by an expert annotator (in argumentation and fallacy theory) that went over both cases of agreement and disagreement to decide the final label. We denote this dataset as CLIMATE where it differs from CLIMATELOGIC (Jin et al., 2022) in three ways: i) it is built using a fallacy scheme specifically developed for misinformation; ii) the fallacious segments are identified by domain experts at `climatefeedback.org` and contain comments which explain fallacious aspects; iii) the segments are mostly 1-3 sentences long.

Final Labels. We unify the labels of similar fallacies (e.g., *False Cause*, *False Causality*, *Causal Oversimplification* → *Causal Oversimplification*). We also rephrase some fallacy types by removing words such as "Appeal to" (e.g,. *Appeal to Emotion* → *Emotional Language*) that tend to throw off generative models causing over prediction of these types as observed in our initial experiments. Some fallacies have partial or full overlap with others across the four schemes. Therefore, we merge these types and use the label of the most frequent or the most representative label of the fallacy type (e.g., *Fallacy of Relevance* → *Red Herring*). We also unify the definitions of fallacy types in prompts across datasets. We end up with 28 unique fallacy types across five datasets ARGOTARIO: 5, PROPAGANDA: 15, LOGIC: 13, COVID-19 and CLIMATE: 9. Complete list of fallacy labels and definitions for all types is shown in Appendix B.

3 Multitask Instruction-based Prompting

Recently, Wei et al. (2022); Sanh et al. (2022) leverage the intuition that NLP tasks can be described via natural language instructions, such as "*Is the sentiment of this movie review positive or negative?*" or "*Translate 'how are you' into Chinese.*". They then take a pre-trained language model and perform *instruction tuning* — fine-tuning the model on several NLP datasets expressed via natural language instructions. Such an approach has several benefits, with the most important one being able to have a unified model for several tasks. Finally, training tasks spanning diverse datasets in a massively multitask fashion improves inference time performance especially for smaller datasets.

Following the success of multitask instruction-based prompting we approach different formulations of fallacies across datasets as different tasks with a generic prompting framework in a single model. We use T5 (Raffel et al., 2020) as the backbone model for training on all five fallacy datasets that have different number and types of fallacies. We hypothesize that when a model is able to learn to recognize fallacy types from multiple datasets, it is more likely able to learn generic traits of fallacy types rather than learning characteristics specific to a single dataset.

A sample list of instructions for each dataset is shown in Figure 1 (Full list in Appendix C). All instructions start with an n-gram (e.g. *'Given a text segment'*) followed by a list of fallacy types with or without their definitions. The complete set of fallacies and definitions are shown in Appendix B. The final component of the instruction is specific to each dataset (e.g., question-answer pair for ARGOTARIO, sentence-fragment or sentence only for PROPAGANDA). The generation target during training and test is one of the fallacies types

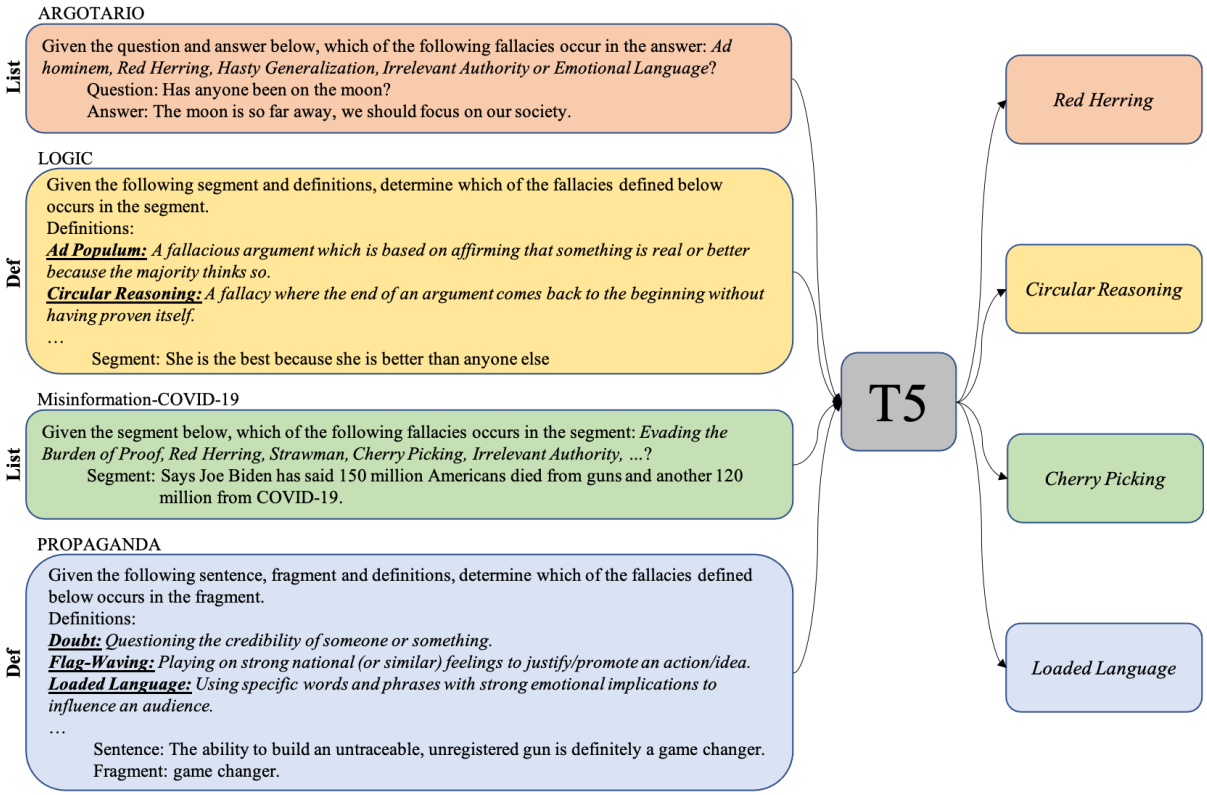

Figure 1: Model and Prompts. **Def**: fallacy definitions in the prompt. **List**: fallacy names listed in the prompt.

that are permissible for each dataset. In addition, we ask the model to generate the fragment that contains the fallacy (PROPAGANDA dataset only) during training to increase the diversity of prompts and instructions during training. Since the overall objective of this work is to have a generic classifier for fallacy and to compare with other classification methods, evaluating the model's ability to correctly generate the fallacious fragment is beyond the scope of this paper. During inference time, we use greedy decoding and select the generated target as the prediction of fallacy type. The evaluation is done using strict string match with the gold fallacy. Model hyperparameters are shown in Appendix A.

4 Evaluation Setup and Results

Given the high imbalance nature of all fallacy datasets, we report both *accuracy* (equivalent to micro F1 as we do not include multi-label instances) and *Macro F1*.

Baselines. We consider the following three models as our baselines: i) zero-shot classification using UnifiedQA (Khashabi et al., 2020); ii) few-shot instruction-tuning of GPT-3 (Brown et al., 2020); and iii) full-shot fine-tuning of BERT (Devlin et al., 2019). UnifiedQA is a question-answering model that is trained on 20 question-answering datasets in different formats and showed generalization capability to unseen data. We use its recent version UnifiedQA-v2 (3B size) (Khashabi et al., 2022) to test the ability of such model to recognize fallacies in zero-shot settings. We also do few-shot instruction-tuning of GPT-3 as many fallacy datasets are of small size, which poses the need for models that can perform well using few-shot training. We setup the instructions in a similar fashion to the ones used for T5 (i.e. *List* prompt in Figure 1). Additionally, we setup instructions with explanations where each few-shot example has a text segment, a fallacy label, and a sentence explaining why the fallacy label is suitable for the text, which is shown to improve the results of few-shot learning (Lampinen et al., 2022).[2] Constrained by the length allowed in the prompt, we use 2-shots per the five fallacy types for the ARGOTARIO dataset, and 1-shot per the nine-to-fifteen fallacy types for the other datasets. Given the high number of fallacy types, it is not feasible to instruction-tune GPT-3 on the 28 unique fallacy types that exist in all five datasets combined. Finally, we fine-tune BERT for 3 epochs on each dataset separately to test its

[2]Example instructions shown in Appendix C

ability to do fallacy recognition. All model hyperparameter details are shown in Appendix A.

We also use a T5-large model trained on each dataset separately using the instructions shown in Figure 1 as a baseline in order to compare the results of single-dataset with multi-dataset training.

4.1 Multitask Instruction-based Prompting vs. Baselines

Baseline Results Looking at the results shown in Table 3, UnifiedQA struggles to have any meaningful results and mostly predicts one or two fallacy types for all examples, which shows the infeasibility for models to perform well in zero-shot settings on a complex task such as fallacy recognition. GPT-3 is able to perform well on ARGOTARIO, even when trained with 1-shot per class, but struggles to beat any full-shot model on the other datasets, which highlights the difficulty of this task for few-shot training. Adding the explanations does not improve the performance, which could have been outweighed by the low number of shots per class and high number of fallacy classes. We notice that BERT has an acceptable performance on the ARGOTARIO dataset (Acc. 44% and F1 38%) that has the lowest number of classes (5 fallacy types), which is also the most balanced dataset compared to the other ones. However, when the number of fallacy classes increases to 9 or more, BERT struggles to have a good performance in any of the two evaluation metrics.

The T5-large models is also trained on each dataset separately using the instructions shown in Figure 1. It has a surprisingly low performance on the ARGOTARIO dataset (Acc. 25% and F1 14%) that is significantly lower than BERT and GPT-3. However, it is able to learn better for datasets with high number of classes (13-15 class) and large training data (e.g. PROPAGANDA and LOGIC).

Multitask Instruction-based Prompting Results We train two sizes of the T5 models (large and 3B) on all datasets combined using the instructions mentioned in Figure 1. This increases the performance significantly on all datasets of the T5-large model compared to its performance when trained on one dataset at a time as shown in Table 3. The numbers further improve when we increase the size of the model from T5-large to T5-3B. This shows the benefit of our unified model based on multitask instruction-based prompting (multi-dataset) for fallacy recognition where we have limited resources and some very small datasets, and also shows the ability of larger models to generalize to the five test sets. The two multi-dataset models always have the best or second best results on all datasets. Also, the T5-3B model is better than T5-large in all accuracy and F1 scores for all datasets excepts accuracy scores for the COVID-19 and CLIMATE where the T5-large is better which could be due to having more correct predictions in the majority classes as the T5-3B is still better in macro F1 scores. To further understand the effect of the model size and prompt choice, we discuss in the next section the per-class performance of four different T5 models.

4.2 Performance of our Unified Model on Fallacy Types

We show the per-class (fallacy type) results of our unified model (multitask instruction-based prompting) using two model sizes (T5-large and T5-3B) and three prompts choices (Def, List, and All) in Tables 4-a to 4-e.

Model Size In general, increasing the model size (from T5-large to T5-3B both trained on all prompts) improves the overall results (especially macro F1) on all datasets. We notice the importance of model size in most datasets for fallacies types that have diversion moves (e.g. *Red Herring* in all datasets, *Strawman* in COVID-19 and CLIMATE, *Whataboutism* in PROPAGANDA) where additional context is usually needed to make accurate predictions. A model with more parameters is in principle better in capturing more information during pretraining, which could be more useful for such fallacies that require more information beyond the provided segment. However, this is not always true for other fallacies of diversion. The results are the same or marginally different for the two model sizes when the fallacy is among the majority training classes (e.g. *Cherry Picking* in COVID-19 and CLIMATE), or inconsistent due to different conceptualizations of a single fallacy across datasets (e.g. *Irrelevant Authority*, more discussion at the end of this Section). Interestingly, the smaller size model (T5-large) has similar performance to the larger model (T5-3B) on some fallacy types with strong lexical cues contained in the text segment (e.g. *Loaded Language*, *Name Calling* and *Slogans* in PROPAGANDA; *Ad Hominem* and *Emotional Language* in LOGIC and ARGOTARIO).

Prompt Choice We also fix the model size (T5-3B) but change the prompts used for training to

Training Data	Shot	Model	Argotario		Propaganda		Logic		Covid-19		Climate	
			Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
–	Zero	UnifiedQA	23	14	04	01	21	08	14	07	08	02
Single	Few	GPT-3	45	39	19	13	20	22	14	09	11	04
	Few+Exp	GPT-3	47	39	13	10	19	22	10	08	10	03
	Full	BERT	44	38	50	25	35	31	25	08	23	04
	Full	T5-Large	25	14	66	30	56	45	26	09	23	04
Multi	Full	T5-Large	59	59	70	41	68	62	**31**	26	**27**	17
	Full	T5-3B	**64**	**64**	**73**	**56**	**70**	**66**	29	**28**	25	**20**

Table 3: Accuracy and Macro F1 scores on all datasets. **Exp**: explanations added to the few shot examples. Numbers in **Bold** represents the best score for each dataset, and underlined numbers are the second best.

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Black-and-White Fallacy	35	32	21	28
Causal Oversimplification	24	48	24	24
Doubt	66	69	61	60
Exaggeration or Minimization	51	61	42	37
Fear or Prejudice	44	56	45	44
Flag-Waving	58	71	73	66
Irrelevant Authority	52	49	26	36
Loaded Language	82	82	80	79
Name Calling or Labeling	83	83	82	82
Red Herring	0	50	18	0
Reductio Ad Hitlerum	0	37	0	0
Slogans	50	51	42	48
Strawman	0	11	0	0
Thought-Terminating Cliches	29	44	38	24
Whataboutism	0	44	43	17
Accuracy	70	73	69	67
Macro F1	41	56	43	39

(a) Propaganda

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Ad Hominem	82	89	84	80
Ad Populum	82	86	83	80
Black-and-White Fallacy	88	84	87	89
Causal Oversimplification	70	81	65	79
Circular Reasoning	59	77	73	71
Deductive Fallacy	53	53	42	46
Emotional Language	71	68	60	57
Equivocation	29	29	29	12
Fallacy of Extension	55	51	62	18
Hasty Generalization	74	70	69	68
Intentional Fallacy	26	33	24	12
Irrelevant Authority	60	70	66	58
Red Herring	60	61	56	47
Accuracy	68	70	67	63
Macro F1	62	66	62	55

(b) Logic

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Ad Hominem	68	68	68	59
Emotional Language	67	68	68	67
Hasty Generalization	41	58	45	54
Irrelevant Authority	75	77	78	71
Red Herring	44	52	41	48
Accuracy	59	64	60	59
Macro F1	59	64	60	59

(c) Argotario

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Causal Oversimplification	29	50	44	42
Cherry Picking	31	31	35	36
Evading the Burden of Proof	47	36	27	41
False Analogy	40	33	50	33
Hasty Generalization	21	0	19	19
Irrelevant Authority	57	24	0	15
Red Herring	0	19	0	0
Strawman	0	24	0	0
Vagueness	8	31	21	0
Accuracy	31	29	28	29
Macro F1	26	28	22	21

(d) Covid-19

Model	T5-L	T5-3B		
Prompt	All	All	Def	List
Causal Oversimplification	37	29	53	32
Cherry Picking	39	41	43	41
Evading the Burden of Proof	0	0	0	0
False Analogy	25	0	0	25
Hasty Generalization	0	0	0	0
Irrelevant Authority	30	27	25	25
Red Herring	0	6	12	11
Strawman	0	46	0	25
Vagueness	22	34	26	34
Accuracy	27	25	29	28
Macro F1	17	20	18	21

(e) Climate

Table 4: F1 scores for each fallacy type for two T5 model sizes (T5-Large and T5-3Billion), and for three prompt choices (**Def**: fallacy definitions in prompt; **List**: fallacy types listed in prompt; **All**: both Def and List prompts) to study the effect of model size and prompt choice. All models are trained on all five datasets combined.

see which prompt is more useful for this task. We mainly experiment with two prompts that include either the definitions of all fallacies or only listing the names of all fallacies. In both cases, the prompt starts with an instruction followed by either definitions or fallacy names then ending with the segment that has the fallacious text. Including both prompts for each training instance yields the best results in most cases as we would expect. However, it seems that some fallacies benefit more from including the definitions in the prompt than others. In general, including the definitions (T5-3B-Def) rather than just fallacy names (T5-3B-List) has higher accuracy and macro F1 scores in 4 out of 5 datasets as shown in Table 4 (exceptions are accuracy in COVID-19 and F1 in CLIMATE). In particular, it seems that definitions are more useful for fallacies that are closely related to other fallacies in one scheme where the definition helps in further clarifying the difference between the two. For example, in PROPAGANDA (Table 4-a) *Thought-Terminating Cliches* are defined as "*words or phrases that offer short, simple and generic solutions to problems*" which is mostly confused with *Loaded Language* by most models, especially ones not trained with definitions. Also in PROPAGANDA, T5-3B-Def has a much higher score than T5-3B-List on *Whataboutism*, which is "*a discrediting technique that accuse others of hypocrisy*" which includes introducing questions about other irrelevant matters. This could have caused models to confuse it with the *Doubt* fallacy.

Fallacy Types Across Datasets There are two fallacies that exist in all five datasets (i.e. *Irrelevant Authority* and *Red Herring*) and two other fallacies that exist in four datasets (i.e. *Causal Oversimplification* and *Hasty Generalization*). We closely look at these fallacies to understand the challenges posed by changes in domain, genre, and annotation guidelines.

Consider the results shown in Tables 4 (a-e) for *Irrelevant Authority*, we can notice three observations: i) T5-large is the best in PROPAGANDA, COVID-19, and CLIMATE; ii) T5-3B-All is the best in LOGIC and marginally second best (to T5-3B-Def) in ARGOTARIO; iii) similar to model size, including the definition in the prompt has inconclusive benefit across datasets. This can be mainly attributed to inconsistency in how this fallacy is defined in different schemes as for example it strictly refers to "*mention of false authority on a given matter*" in COVID-19, while it additionally includes "*referral to a valid authority but without supporting evidence*" in PROPAGANDA (all definitions provided in Appendix B).

Similarly, no single model is consistently better in detecting *Red Herring* across all datasets as shown in Tables 4 (a-e). This, however, is more likely caused by the different format this particular fallacy has in different domains and genres as it consists of shorter phrases in PROPAGANDA, asking irrelevant or misleading questions in CLIMATE, and mentions of irrelevant entities in LOGIC.

Causal Oversimplification has more consistent results as shown in Tables 4 (a,b,d,e) where the T5-3B-All model has the best results in three out of four datasets. This illustrates that while the notion of this fallacy might differ across datasets, it still strongly shares common generic features (e.g. the existence of a causal relation) that make it distinguishable by a single model in different settings.

Finally, the results for *Hasty Generalization* shown in Tables 4 (b-e) indicate that detecting this fallacy becomes more challenging when other similar fallacies exist in a fallacy scheme (e.g. *Cherry Picking* in COVID-19 and CLIMATE), and less challenging when other fallacies in the scheme are further away (e.g. LOGIC and ARGOTARIO).

Nevertheless, this multitask setup provides the model with the opportunity to learn to detect specific fallacy types as they are expressed differently, and grouped with different fallacies, which consistently and significantly improves the overall results of fallacy recognition over single-scheme (or single dataset) models.

4.3 Error Analysis

In order to better understand model errors and quality of annotations for this complex task for both humans and machines, an expert looked at 70 wrongly predicted examples from the PROPAGANDA datasets (5 examples each from 14 propaganda technique, *Strawman* was not included due to low counts). First, the expert looked only at the sentence and the fragment identified by the gold annotation as containing a fallacy and she independently annotated the propaganda technique at stake. Comparing this annotation with gold labels and model prediction (T5-3B-All), it turns out that the expert annotator agreed with the gold label in 75% of the cases, and with the model prediction

His opinion is: "She may very well believe everything she's saying, and that is one of the signs of lunacy, believing something that isn't real." And her lawyer is even loonier	Doubt Name Calling or Labeling Doubt
"Christianity is Europe's last hope," Orban told an audience of party faithful at the foot of the Royal Castle in Budapest.	Slogans Flag-Waving Flag-Waving
"Orban is openly Christian and seems to understand something that many do not and that is you do not allow a wholesale flood of antichrists to pour into your country."	Flag-Waving Name Calling or Labeling Flag-Waving

Table 5: Example sentences from PROPAGANDA with gold label, model prediction and expert annotation. Underlined text highlights the propagandistic fragment.

in 15%, while she chose a different label in 10% of the cases. Table 5 shows three examples along with gold labels, model predictions, and expert annotations.

Consider the first example in Table 5 that has *Doubt* as the gold label. The expert agrees that the propaganda technique used rests on questioning the credibility of the lawyer (*Doubt*), even though the adjective "lunatic" is a literal instance of *Name Calling*. Thus, the label predicted by the model is not wrong, but less relevant since it is the lack of trustworthiness the most effective feature in undermining the antagonist's stance, regardless whether it is due to lunacy or lack of integrity.

In the second example of Table 5, the expert agrees with the model prediction of a *Flag-Waving* fallacy in the underlined segment rather than a *Slogan* as the gold label. The term "*last hope*" can be considered a slogan, however, when we consider the full propagandistic segment that includes the word "*Christianity*", it maps better to *Flag-Waving* as it has been defined in the guidelines (and included in the prompt) as "*Playing on strong national feeling (or to any group)...*".

The third example highlights even more the importance of the selected fragment in the prompt: without considering the reference to the "*antichrist*" threat, it is not possible to understand that the sentence is playing on a religious-based national feeling.

Considering the analysis of the 70 examples in the PROPAGANDA dataset, the following general observations are found: i) some fallacious segments can map to more than one fallacy, especially when one of the two is a language fallacy (e.g., *Name Calling, Exaggeration, Loaded Language*). In such cases, the model tends to privilege the language fallacy type, even if usually not the most relevant from an argumentative perspective; ii) for some cases, the expert annotator had to read more context beyond the sentence; iii) for some cases, the expert agreed with the gold label but disagreed with the boundaries of the annotated fragment by choosing a larger or more informative one.

In light of this, improving automatic fallacy identification may entail i) considering additional context; ii) adopting a fallacy scheme with a heuristics that imposes an order into fallacy recognition (structural fallacy followed by diversion and logical fallacies with language fallacies at last when all the others are excluded).

5 Related Work

Fallacy There are various typologies of fallacies that address informal logic traditions or rules of ideal critical discussion (Hansen, 1996; Van Eemeren et al., 2002; Tindale, 2007; Walton et al., 2008; Damer, 2012). This intersects with propaganda techniques that focus on faulty reasoning and emotional appeals to accomplish persuasion (Miller, 1939; Jowett and O'Donnell, 2012; Torok, 2015; Weston, 2018). Computational work on fallacy recognition includes fallacies in dialogue (Habernal et al., 2017; Sheng et al., 2021), argument sufficiency (Stab and Gurevych, 2017), name calling on Reddit (Habernal et al., 2018), non-sequitur fallacy in legal text (Nakpih and Santini, 2020), logical fallacies (Jin et al., 2022), fallacies in misinformation (Musi et al., 2022), as well as propaganda techniques (Da San Martino et al., 2019b). However, all previous work is limited to one fallacy scheme while we develop a model that detects fallacies in five datasets across four fallacy schemes.

Prompting Using prompts has emerged as a generic framework to train natural language processing models on multiple tasks using prefix text (Raffel et al., 2020), and few-shot prompt-tuning of GPT-3 (Brown et al., 2020). This was followed by multiple studies that use prompts on smaller size models using few and full shots on tasks such as natural language inference (Schick and Schütze, 2021b), text classification (Schick and Schütze, 2021a; Gao et al., 2021), relation extraction (Chen et al., 2022), and using instruction prompts for mul-

tiple tasks (Mishra et al., 2022; Sanh et al., 2022). We follow a similar setup by training a T5 model using instruction prompts for different formulations of fallacy recognition approached as multiple tasks.

6 Conclusion

We introduced a unified model using multitask instruction-based prompting for solving the challenges faced by the fallacy recognition task. We could unify all the datasets by converting 28 fallacy types across 5 different datasets into natural language instructions. We showed that our unified model is better than training on a single dataset. We analyzed the effect of model size and prompt choice on the detection of specific fallacy types that could require additional knowledge better captured by bigger models (e.g., diversion fallacies such as *Red Herring*), and the distinction between similar fallacies better detected by more comprehensive prompts that include definitions of fallacy types (e.g., *Doubt* vs. *Whataboutism*). We analyzed the differences of fallacy types that appear in multiple fallacy schemes across the five datasets and showed that one fallacy type could have multiple meanings which further increases the complexity of this task (e.g., *Irrelevant Authority*). We conducted a thorough error analysis and released a new fallacy dataset for fact-checked content in the climate change domain.

Limitations

In the current setup, we consider all examples as fallacious or partially fallacious and do not include a "No Fallacy" class, which some of the fallacy datasets have. Based on this assumption, the model's task is to detect the type of fallacy given a fallacious example. Including "No Fallacy" makes the datasets severely imbalanced (e.g. 70% of PROPAGANDA and 50% of COVID-19 are labeled as "No Fallacy"). We elected to remove it for this work since not all datasets have a "No Fallacy" class (e.g. LOGIC) and since this class is bigger than all 28 fallacy class combined. Even with our initial experiments with downsampling of "No Fallacy" using BERT, the results were not promising. This setup is in line with the propaganda technique classification task (Da San Martino et al., 2020) and the logical fallacy detection task (Jin et al., 2022) that all do not include "No Fallacy" class. We leave further experimentation of pipeline or joined approaches to separate fallacies from non-fallacies text for future work. Other limitations include the need for external knowledge and the multi-labeling nature of some examples as discussed in Section 4.3, which we leave for future work.

We experiment with the second and third largest sizes of the T5 model, T5-3B (11GB) and T5-large (3GB) and do not run experiments with T5-11B (40GB) due to lack of resources. The T5-3B is run on 2 Nvidia A-100 GPUs with 40GB memory each with a batch size of 2. These GPU requirements could pose a limitation on using such models in resource-poor settings. They could also have environmental impacts if trained (and re-trained) for longer periods of time. The training time of the T5-3B on 2 GPUs for 5 epochs is on average 2-3 hours depending on the size of the dataset.

Ethics Statement

The intended use of our fallacy recognition model is as a human-in-the-loop assistant to signal misinformation or for teaching humans how to identify and avoid using fallacies.

Acknowledgments

The first author is grateful for the KACST Graduate Studies Scholarship that supported his PhD research for five years.

References

Tariq Alhindi, Jonas Pfeiffer, and Smaranda Muresan. 2019. Fine-tuned neural models for propaganda detection at the sentence and fragment levels. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 98–102, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth*

Workshop on Semantic Evaluation, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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. 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.

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. 2012. *Attacking faulty reasoning*. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

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.

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.

Charles L Hamblin. 2022. Fallacies. Advanced Reasoning Forum.

Hans V Hansen. 1996. Aristotle, whately, and the taxonomy of fallacies. In *International Conference on Formal and Applied Practical Reasoning*, pages 318–330. Springer.

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

Garth S Jowett and Victoria O'Donnell. 2012. What is propaganda, and how does it differ from persuasion. *Propaganda & persuasion*, pages 1–48.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Clyde R. Miller. 1939. *The Techniques of Propaganda. From "How to Detect and Analyze Propaganda,"*. an address given at Town Hall. The Center for learning.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

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*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

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.

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

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming.

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

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.

A Model Hyperparameters

We use huggingface's implementation (Wolf et al., 2020) of the T5 model (large and 3B) where we train all models for 5 epochs choosing the epoch with lowest evaluation loss as the final model. The models are run with 1e-4 learning rate, Adam optimizer, batch size 2, gradient accumulation steps 512, maximum source length 1024, maximum target length 64. At inference time, the target is generated using greedy decoding (beam search of size 1) with no sampling and default settings for T5. The generated target is then compared with the fallacies in the given scheme and the prediction is counted as correct if they are the same using strict string match.

We also use huggingface's implementation of BERT (base) and fine-tune the model for 3 epochs with 1e-5 learning rate, batch size 16, maximum sequence length 256.

For GPT-3, we use the completion API of OpenAI (Brown et al., 2020) using their large engine that is trained with instructions (*text-davinci-002*) with temperature 0, max generated tokens 150 and other parameters kept at default value (e.g. top_p=1). The generated target is considered correct if it has the gold fallacy (even with additional text). Since GPT-3 is trained with few-shots only, it sometimes generates some generic prefix, repeats the text segment, or generates more than one fallacy.

B Fallacy Datasets

We list in Tables 6 and 7 all the definitions and fallacy labels used in our prompts. As mentioned in Section 2, we unify the definitions and labels for fallacies that fully or partially overlap. Additionally, in the same tables we show the original labels and definitions for all four fallacy schemes as they are released by (Habernal et al., 2017) for ARGOTARIO, (Da San Martino et al., 2019b) for PROPAGANDA, (Jin et al., 2022) for LOGIC, and (Musi et al., 2022) for MISINFORMATION that is used by the COVID-19 and CLIMATE datasets. We also show counts of fallacy types in training/dev/test splits for all datasets in Table 8. Below is a detailed description of the four existing fallacy datasets.

ARGOTARIO Introduced by Habernal et al. (2017), the Argotario dataset consists of five fallacies in dialogue between players in game settings. The five fallacy types are: *Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization, irrelevant authority*, in addition to the *No Fallacy* type. These types are selected because they are: common in argumentative discourse, distinguishable from each other, and have different difficulty levels. Players in the game are presented with a topic (question), which they answer using one of the fallacy types. Other players then try to predict the fallacy type written by author of the answer. The final label is determined when at least four players agree with the author of the answer on the type of fallacy. Each instance consist of a question-answer pair and one out of five fallacy labels.

PROPAGANDA Da San Martino et al. (2019b) identified 18 propaganda techniques that appear in news articles. We focus on the following 15 of them that have a fallacy and frequent enough in the data: *Loaded Language, Name Calling or Labeling, Exaggeration or Minimization, Doubt, Appeal to Fear/Prejudice, Flag-Waving, Causal Oversimplification, Slogans, Appeal to Authority, Black-and-White Fallacy, Thought-Terminating Cliche, Whataboutism, Reductio ad Hitlerum, Red Herring,* and *Strawman*. We ignore propaganda techniques that do not have an argumentative fallacy (e.g. *Repetition*) or not frequent enough in the data (e.g. *Bandwagon, OIVC*). The authors annotate the text spans and propaganda technique (fallacy type) in 451 articles from 48 news outlets allowing multiple labels and partial overlap of text spans. We frame this at the sentence level where the fallacy type becomes the label of the sentence if the fragment is included within the sentence. For sentences with multiple fragments, we consider the label of the longer fragment. We ignore propaganda fragments that span across multiple sentences. This is the biggest dataset in our experiments but it is also the most imbalance one where 6 out the 18 propaganda techniques represent more than 80% of all propagandistic segments. Each training instance consists of a sentence, a fragment, and one out of fifteen fallacy labels.

LOGIC Jin et al. (2022) collected examples of logical fallacies from educational websites on fallacies such as Quizziz, study.com and ProProfs. They identified 13 types of fallacies in the dataset using Wikipedia[3] as a reference. The fallacy types are: *Faulty Generalization, False Causality, Circular Claim, Ad Populum, Ad Hominem, Deductive*

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

	Fallacy Type	Definition
(Habernal et al., 2017)	Ad Hominem	**The opponent attacks a person instead of arguing against the claims that the person has put forward.**
	Appeal to Emotion (Emotional Language)	**This fallacy tries to arouse non-rational sentiments within the intended audience in order to persuade.**
	Hasty Generalization	The argument uses a sample which is too small, or follows falsely from a sub-part to a composite or the other way round.
	Irrelevant Authority	While the use of authorities in argumentative discourse is not fallacious inherently, appealing to authority can be fallacious if the authority is irrelevant to the discussed subject.
	Red Herring	This argument distracts attention to irrelevant issues away from the thesis which is supposed to be discussed.
(Da San Martino et al., 2019b)	Black and White Fallacy	**Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an the extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship).**
	Causal Oversimplification	**Assuming a single cause or reason when there are actually multiple causes for an issue.**
	Doubt	**Questioning the credibility of someone or something.**
	Exaggeration or Minimization	**Either representing something in an excessive manner: making things larger, better, worse or making something seem less important than it really is**
	Appeal to fear/prejudice (Fear or Prejudice)	**Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases the support is based on preconceived judgements.**
	Flag-Waving	**Playing on strong national feeling (or to any group) to justify/promote an action/idea.**
	Appeal to Authority (Irrelevant Authority)	Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered. We consider the special case in which the reference is not an authority or an expert in this technique, although it is referred to as Testimonial in literature.
	Loaded Language	**Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.**
	Name Calling or Labeling	**Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises.**
	Red Herring	Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.
	Reductio Ad Hitlerum	**Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.**
	Slogans	**A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.**
	Strawman	**When an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition.**
	Thought-Terminating Cliches	**Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or distract attention away from other lines of thought.**
	Whataboutism	**A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.**
(Jin et al., 2022)	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.
	Ad Populum	**A fallacious argument which is based on affirming that something is real or better because the majority thinks so.**
	False Dilemma (Black and White Fallacy)	A claim presenting only two options or sides when there are many options or sides.
	False Causality (Causal Oversimplification)	A statement that jumps to a conclusion implying a causal relationship without supporting evidence
	Circular Reasoning	**A fallacy where the end of an argument comes back to the beginning without having proven itself.**
	Deductive Fallacy	**An error in the logical structure of an argument.**
	Appeal to Emotion (Emotional Language)	Manipulation of the recipient's emotions in order to win an argument.
	Equivocation	**An argument which uses a phrase in an ambiguous way, with one meaning in one portion of the argument and then another meaning in another portion.**
	Fallacy of Extension	**An argument that attacks an exaggerated/caricatured version of an opponent's.**

Table 6: Fallacy Names and Definitions (**Bold**: definition of this fallacy used in all prompts for across datasets)

	Faulty Generalization (Hasty 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.
	Intentional Fallacy	**Some intentional/subconscious action/choice to incorrectly support an argument.**
	Fallacy of Credibility (Irrelevant Authority)	An appeal is made to some form of ethics, authority, or credibility.
	Fallacy of Relevance (Red Herring)	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).
(Musi et al., 2022)	Evading Burden of Proof	**A position is advanced without any support as if it was self-evident.**
	Cherry Picking	**The act of choosing among competing evidence that which supports a given position, ignoring or dismissing findings which do not support it.**
	Red Herring	**The argument supporting the claim diverges the attention to issues which are irrelevant for the claim at hand.**
	Strawman	The arguer misinterprets an opponent's argument for the purpose of more easily attacking it, demolishes the misinterpreted argument, and then proceeds to conclude that the opponent's real argument has been demolished.
	False Authority (Irrelevant Authority)	**An appeal to authority is made where the it lacks credibility or knowledge in the discussed matter or the authority is attributed a tweaked statement.**
	Hasty Generalization	**A generalization is drawn from a sample which is too small, not representative or not applicable to the situation if all the variables are taken into account.**
	False Cause (Causal Oversimplification)	X is identified as the cause of Y when another factor Z causes both X and Y OR X is considered the cause of Y when actually it is the opposite
	Post Hoc (Causal Oversimplification)	It is assumed that because B happens after A, it happens because of A. In other words a causal relation is attributed where, instead, a simple correlation is at stake
	False Analogy	**because two things [or situations] are alike in one or more respects, they are necessarily alike in some other respect.**
	Vagueness	**A word/a concept or a sentence structure which are ambiguous are shifted in meaning in the process of arguing or are left vague being potentially subject to skewed interpretations.**

Table 7: (*continue*) Fallacy Names and Definitions (**Bold**: definition of fallacy used in all prompts across datasets)

	Fallacy	ARGOTARIO train/dev/test	PROPAGANDA train/dev/test	LOGIC train/dev/test	COVID-19 train/dev/test	CLIMATE train/dev/test	Total train/dev/test	Total All
1	Ad Hominem	102 /26/ 31	–/–/–	406 /64/ 81	–/–/–	–/–/–	508 /90/ 112	710
2	Ad Populum	–/–/–	–/–/–	296 /81/ 62	–/–/–	–/–/–	296 /81/ 62	439
3	B&W Fallacy	–/–/–	60 /16/ 19	192 /40/ 25	–/–/–	–/–/–	252 /56/ 44	352
4	Causal Ov.simp.	–/–/–	111 /28/ 34	303 /49/ 36	36 /10/ 10	39 /10/ 11	489 /97/ 91	677
5	Cherry Picking	–/–/–	–/–/–	–/–/–	76 /20/ 23	67 /17/ 21	143 /37/ 44	224
6	Circular Reason.	–/–/–	–/–/–	238 /40/ 35	–/–/–	–/–/–	238 /40/ 35	313
7	Deductive	–/–/–	–/–/–	205 /28/ 31	–/–/–	–/–/–	205 /28/ 31	264
8	Doubt	–/–/–	263 /66/ 82	–/–/–	–/–/–	–/–/–	263 /66/ 82	411
9	Emotional Lang.	150 /38/ 47	–/–/–	230 /38/ 41	–/–/–	–/–/–	380 /76/ 88	544
10	Equivocation	–/–/–	–/–/–	62 /13/ 11	–/–/–	–/–/–	62 /13/ 11	86
11	Evad Burd Prf	–/–/–	–/–/–	–/–/–	76 /20/ 23	31 /8/ 9	107 /28/ 32	167
12	Exag/Mini	–/–/–	304 /76/ 94	–/–/–	–/–/–	–/–/–	304 /76/ 94	474
13	Extension	–/–/–	–/–/–	187 /31/ 46	–/–/–	–/–/–	187 /31/ 46	264
14	False Analogy	–/–/–	–/–/–	–/–/–	13 /5/ 3	17 /5/ 5	30 /10/ 8	48
15	Fear/Prejudice	–/–/–	131 /33/ 41	–/–/–	–/–/–	–/–/–	131 /33/ 41	205
16	Flag-Waving	–/–/–	145 /37/ 45	–/–/–	–/–/–	–/–/–	145 /37/ 45	227
17	Hasty General.	104 /26/ 32	–/–/–	561 /128/ 123	54 /15/ 16	4 /2/ 2	723 /171/ 173	1,067
18	Intentional Fal.	–/–/–	–/–/–	215 /34/ 26	–/–/–	–/–/–	215 /34/ 26	275
19	Irrelevant Auth.	92 /24/ 29	57 /15/ 17	196 /18/ 33	26 /8/ 8	32 /8/ 10	403 /73/ 97	573
20	Loaded Lang.	–/–/–	1,331/333/416	–/–/–	–/–/–	–/–/–	1,331/333/416	2,080
21	Name Calling	–/–/–	685 /172/ 214	–/–/–	–/–/–	–/–/–	685 /172/ 214	1,071
22	Red Herring	115 /29/ 35	16 /4/ 10	214 /43/ 46	28 /8/ 8	44 /12/ 13	417 /96/ 112	625
23	Reductio AH.	–/–/–	33 /9/ 10	–/–/–	–/–/–	–/–/–	33 /9/ 10	52
24	Slogans	–/–/–	84 /22/ 26	–/–/–	–/–/–	–/–/–	84 /22/ 26	132
25	Strawman	–/–/–	4/1/6	–/–/–	28 /8/ 8	23 /6/ 7	55 /15/ 21	91
26	Thought-Term.	–/–/–	48 /12/ 14	–/–/–	–/–/–	–/–/–	48 /12/ 14	74
27	Vagueness	–/–/–	–/–/–	–/–/–	53 /15/ 23	48 /12/ 14	101 /27/ 37	165
28	Whataboutism	–/–/–	33 /9/ 10	–/–/–	–/–/–	–/–/–	33 /9/ 10	52
	Total (tr/de/te)	563 /143/ 174	3,305 /833/ 1,038	3,305 /607/ 596	390 /109/ 122	305 /80/ 92	7,868 /1,772/ 2,022	
	Total (All)	880	5,176	4,508	621	477	11,662	

Table 8: Counts of fallacy types in each split across all datasets.

Fallacy, Appeal to Emotion, False Dilemma, Equivocation, Fallacy of Extension, Fallacy of Relevance, Fallacy of Credibility and *Intentional Fallacy*. Each training instance consists of a text segment (e.g. dialogue, sentence) and one of thirteen fallacy labels. The authors also introduce another challenge dataset: CLIMATELOGIC that follows the same fallacy scheme. However, it contains text segments that are too long (e.g. multiple paragraphs) with no annotations of smaller fallacious fragments like the Propaganda dataset. Therefore, CLIMATELOGIC is beyond the scope of this study.

Misinformation Musi et al. (2022) identified 10 fallacies though analysis of fact-checked news (article and social media posts) about COVID-19. They consider fallacies as indicators of misinformation, which they define as misleading news that is not necessarily false communicated with the intention to deceive, thus making it harder to detect and fact-check. The fallacies are: *Structural (Evading the Burden of Proof), Diversion (Cherry Picking, Strawman, Red Herring, False Authority), Logical (Hasty Generalization, Post Hoc, False Cause, False Analogy), and Language (Vagueness)*. They annotate 1,135 covid-19 news and social media posts (621 fallacious) that are fact-checked by five fact-checking organizations.

C Instructions

We list all instructions used during training in Table 10. ARGOTARIO, LOGIC and COVID-19 have two instructions per example: *List*: fallacy types listed in prompt, and *Def* fallacy definitions included in prompt. For PROPAGANDA, since each instance is sentence with a marked fallacious fragment, we construct three *List* instructions where the fragment is included in the first instruction, removed completely in the second instruction (no fragment in Table 10), and moved to the generation target in the third instruction (*Frag*). The same three instructions are done using *Def* prompts making the total six instructions per training example. For CLIMATE, each instance is constructed using four instructions: *List* and *Def* with and without fact-checkers comments (*Com*). These additional instructions for PROPAGANDA and CLIMATE are included during training only to increase the diversity of prompts.

Also, as discussed in 4.1, we use few shot instruction-tuning of GPT-3 with and without explanations. The instructions that do not include explanations follow the same format of the *List* prompts shown in Table 10 where it starts "Given a text segment ..." followed by a list of fallacy types and then the few shot examples that include a text segment and a fallacy type. Additionally, we write explanations after each few-shot example in the instruction prompt which explains why a given text segment is labeled with the fallacy type. The explanations follow the fallacy type labels as shown in Table 9.

Given the question and answer pairs below, which of the following fallacies occur in the answers: *Emotional Language, Red Herring, Hasty Generalization, Ad Hominem*, or *Irrelevant Authority*?

---

1) **Question:** Is Christianity a peaceful religion?
Answer: You are the antichrist, you want to destroy our belief in god.
Fallacy: Ad Hominem
Explanation: It is an ad hominem becase the speaker is attacked for his bad intentions and not for the point she is making.

2) **Question:** Is television an effective tool in building the minds of children?
Answer: All TV-Shows are bad. Look at "the bachelor". Children cannot learn from it.
Fallacy: Hasty Generalization
Explanation: It is a hasty generalization since the evaluation of a whole category is drawn from the evaluation of a single element of the category.

...

5) **Question:** Should we allow animal testing for medical purposes?
Answer: No, animals are so cuuuuteeeeeeee!!!
Fallacy: Emotional Language
Explanation: It is a fallacy of emotional language since the argument appels to positive emotions associated to animals' appearances.

6) Question: Should gorillas be held in zoos
Answer: No, I don't like gorillas.

Fallacy: Red Herring

Table 9: Example of GPT-3 few-shot instruction with explanations. Test Example Generated Fallacy Type

Name	Data	Instruction	Target
List	ARGOTARIO	Given the question and answer below, which of the following fallacies occur in the answer: {fallacies}? Question: {question}Answer: {answer}	{fallacy}
Def	ARGOTARIO	Given the question, answer and definitions, determine which of the fallacies defined below occur in the answer. Definitions: {definitions} Question: {question}Answer: {answer}	{fallacy}
List	PROPAGANDA	Given the sentence and fragment below, which of the following fallacies occur in the fragment: {fallacies}? Sentence: {sentence} Fragment: {fragment}	{fallacy}
List	PROPAGANDA (no fragment)	Given the sentence below, which of the following fallacies does it have: {fallacies}? Sentence: {sentence}	{fallacy}
List +Frag	PROPAGANDA	Given the sentence below, detect the fallacious fragment in it, and the type of fallacy of the fragment from the following: {fallacies}. Sentence: {sentence}	{fragment} {fallacy}
Def	PROPAGANDA	Given the following sentence, fragment and definitions, determine which of the fallacies defined below occurs in the fragment: {fallacies}. Definitions: {definitions} Sentence: {sentence} Fragment: {fragment}	{fallacy}
Def	PROPAGANDA (no fragment)	Given the sentence and definitions below, determine which of the fallacies defined below occurs in the sentence: {fallacies}. Definitions: {definitions} Sentence: {sentence}	{fallacy}
Def +Frag	PROPAGANDA	Given the following sentence and definitions, detect the fallacious fragment in the sentence, and the fallacy type of that fragment from the ones defined below. Definitions: {definitions} Sentence: {sentence}	{fragment} {fallacy}
List	LOGIC	Given the segment below, which of the following fallacies does it have: {fallacies}? Segment: {segment}	{fallacy}
Def	LOGIC	Given the following segment and definitions, determine which of the fallacies defined below occurs in the segment. Definitions: {definitions} Segment: {segment}	{fallacy}
List	COVID-19	Given the segment below, which of the following fallacies does it have: {fallacies}? Segment: {segment}	{fallacy}
Def	COVID-19	Given the following segment and definitions, determine which of the fallacies defined below occurs in the segment. Definitions: {definitions} Segment: {segment}	{fallacy}
List	CLIMATE	Given the segment below, which of the following fallacies does it have: {fallacies}? Segment: {segment}	{fallacy}
Def	CLIMATE	Given the segment and definitions below, determine which of the fallacies defined below occurs in the segment. Definitions: {definitions} Segment: {segment}	{fallacy}
List +Com	CLIMATE	Given the segment and comment below, which of the following fallacies occur in the segment: {fallacies}? Segment: {segment} Comment: {comment}	{fallacy}
Def +Com	CLIMATE	Given the segment, comment and definitions below, determine which of the fallacies defined below occurs in the segment. Definitions: {definitions} Segment: {segment} Comment: {comment}	{fallacy}

Table 10: All Instructions. **Def**: fallacy definition included in the prompt. **List**: fallacy types listed in the prompt. **Frag**: fallacious fragment and fallacy type are the generation targets. **Com**: fact-checker comments are included, which sometimes explain the fallacy of the segment (**Frag** and **Com** are done during training only).