Detecting Fallacies in Climate Misinformation: A Technocognitive Approach to Identifying Misleading Argumentation

Francisco Zanartu, John Cook, Markus Wagner, Julian Garcia

November 14, 2023

Abstract

Misinformation about climate change is a complex problem that requires holistic and interdisciplinary solutions at the intersection between technology and psychology. One proposed solution is a "technocognitive" approach, involving the synthesis of psychological and computer science research. Psychological research has identified that interventions in response to misinformation require both fact-based (e.g., factual explanations) and technique-based (e.g., explanations of misleading techniques) content. However, little progress has been made on documenting and detecting fallacies in climate misinformation. In this study, we apply a previously developed critical thinking methodology for deconstructing climate misinformation, in order to develop a dataset mapping different types of climate misinformation to reasoning fallacies. This dataset is used to train a model to detect fallacies in climate misinformation. The fallacies that are easiest to detect include fake experts and anecdotal arguments. Fallacies that require background knowledge, such as oversimplification, misrepresentation, and slothful induction, are relatively more difficult to detect. This research lays the groundwork for development of solutions where automatically detected climate misinformation can be countered with generative technique-based corrections.

1 Introduction

Misinformation about climate change reduces climate literacy and support for policies that mitigate climate impacts (Ranney and Clark, 2016) while exacerbating public polarization (Cook et al., 2017). Efforts to communicate the reality of climate change can be cancelled out by misinformation (Van der Linden et al., 2017) and ignorance about the strong degree of public acceptance causes ”climate silence” (Geiger and Swim, 2016). These impacts necessitate interventions that neutralize their negative influence.

A growing body of evidence has documented effective ways to reduce the impact of misinformation. Two leading communication approaches are fact-based and technique-based. Fact-based corrections—also described as topic-based (Schmid and Betsch, 2019)—involve exposing how misinformation is false through factual explanations. Technique-based corrections—also described as logic-based (Banas and Miller, 2013; Vraga et al., 2020)—involve explaining misleading rhetorical techniques and logical fallacies used in misinformation. Schmid and Betsch (2019) found that both fact-based and technique-based corrections were effective in countering misinformation. However, Vraga et al. (2020) found that technique-based corrections outperformed fact-based corrections as they were equally effective whether the correction was encountered before or after the misinformation. In contrast, fact-based corrections were ineffective if misinformation was shown afterwards, leading to a cancelling out effect. This result is consistent with other studies finding that factual explanations can be cancelled out if encountered alongside contradicting misinformation (Cook et al., 2017; McCright et al., 2016; Van der Linden et al., 2017). Synthesising the body of psychological research on countering misinformation, the recommended structure of an effective debunking contains both a fact-based element explaining the facts relevant to the misinforming argument and a technique-based element explaining the misleading rhetorical techniques or logical fallacies found in the misinforming argument (Lewandowsky et al., 2020).

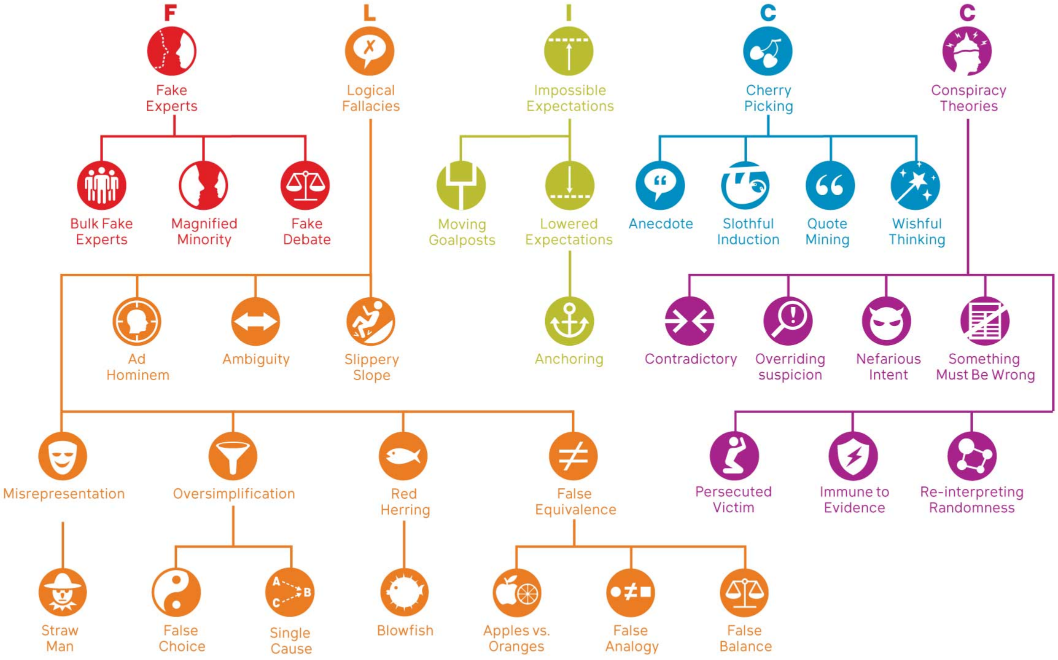


Figure 1: FLICC taxonomy of misinformation techniques and logical fallacies (Cook, 2020).

Increasing research attention has focused attention on understanding and countering the techniques used in misinformation. One framework identifies five techniques of science denial—fake experts, logical fallacies, impossible expectations, cherry picking, and conspiracy theories (Diethelm and McKee, 2009)—summarised with the acronym FLICC. These techniques, found in a range of scientific topics such as climate change, evolution, and vaccination, have been developed into a more comprehensive taxonomy shown in Figure 1 (Cook, 2020). A critical thinking methodology was developed for deconstructing and analysing climate misinformation in order to identify misleading logical fallacies (Cook et al., 2018). This methodology has been applied to contrarian climate claims in order to identify the fallacies used in specific climate myths (Flack et al., 2023). The fallacies identified in climate misinformation, as well as their definitions, are listed in Table 1. The two types of fallacies are structural, where the presence of the fallacy can be gleaned from the structure of the text, and background knowledge, where certain factual knowledge is required in order to perceive that the argument is fallacious. Table 1 also presents the textual structure of each fallacious argument.

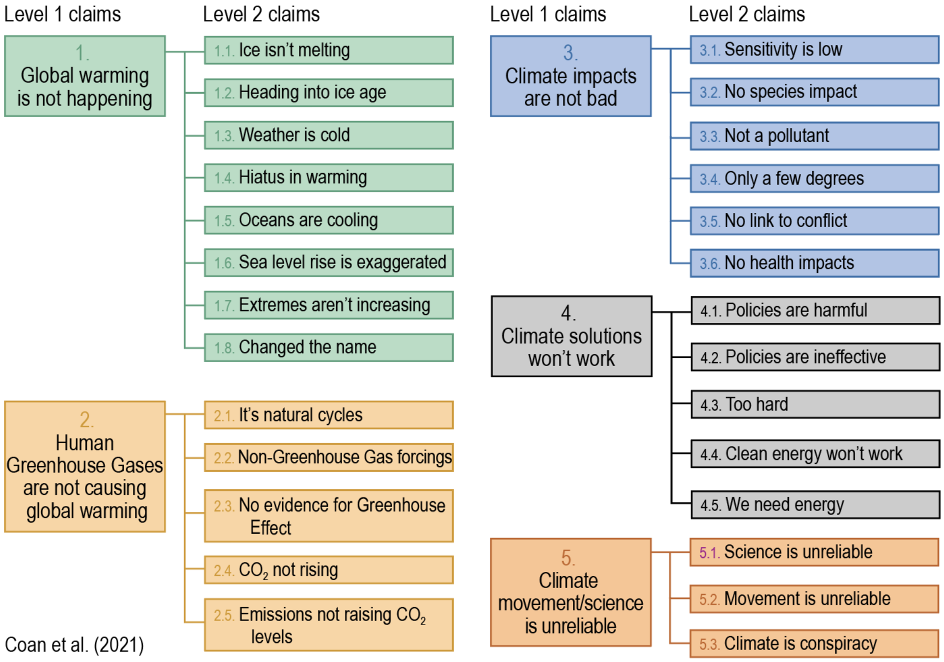
While these theoretical frameworks have been developed based on psychological and critical thinking research, developing practical solutions is challenging for various reasons. Misinformation is perceived by the public as more novel than factual information and consequently spreads faster and farther through social networks than true news (Vosoughi et al., 2018). Once people accept a piece of misinformation, they continue to be influenced by it even if they remember a retraction (Ecker et al., 2010). To address these challenges, research has begun to focus on pre-emptive or rapid response solutions that can reduce the spread and influence of misinformation.

One proposed solution is automatic and instantaneous detection and fact-checking of misinformation, known as the "holy grail of fact-checking" (Hassan et al., 2015). Topic analysis offers the ability to analyse large datasets with unsupervised models that can identify key themes. This approach has been applied to conservative think-tank (CTT) websites, a prolific source of climate misinformation (Boussalis and Coan, 2016). Similarly, topic modelling has been combined with network analysis to find an association between corporate funding and polarizing climate text (Farrell, 2016). Lastly, topic modelling of newspaper articles has been used to identify economic or uncertainty framing about climate change (Stecula and Merkley, 2019). While the unsupervised approach offers general insights about the nature of climate misinformation with large datasets, it doesn't facilitate detection of specific misinformation claims which is necessary in order to generate automated fact-checks.

To address this shortcoming, a supervised machine model—described as the CARDS model (Com-

Fallacy	Type	Definition	Argument Structure
Ad hominem	Structural	Attacking a person/group instead of addressing their arguments	A has a negative trait. Therefore, A is not credible.
Anecdote	Structural	Using personal experience or isolated examples instead of sound arguments or compelling evidence	Y occurred once with X. Therefore, Y will occur every time with X.
Cherry Picking	Structural	Selecting data that appear to confirm one position while ignoring other data that contradicts that position	Group A are lying to us to implement a secret plan.
Conspiracy theory	Structural	Proposing that a secret plan exists to implement a nefarious scheme such as hiding a truth	A is true. B is why the truth cannot be proven. Therefore, A is true.
Fake experts	Structural	Presenting an unqualified person or institution as a source of credible information.	
False choice	Structural	Presenting two options as the only possibilities, when other possibilities exist	P or Q. P. Therefore, not Q.
False equivalence	Structural	Incorrectly claiming that two things are equivalent, despite the fact that there are notable differences between them.	A and B both share characteristic C. Therefore, A and B share some other characteristic D.
Impossible expectations	Structural	Demanding unrealistic standards of certainty before acting on the science	There is not enough data or research about X to understand X properly.
Misrepresentation	Background knowledge	Misrepresenting a situation or an opponent's position in such a way as to distort understanding	
Oversimplification	Background knowledge	Simplifying a situation in such a way as to distort understanding, leading to erroneous conclusions	
Single cause	Structural	Assuming a single cause or reason when there might be multiple causes or reasons	X caused Y; therefore, X was the only cause of Y.
Slothful induction	Background knowledge	Ignoring relevant evidence when coming to a conclusion	

Table 1: Fallacy types, definitions, and argument structure.


Level 1 claims
Level 2 claims
1. Global warming is not happening
1.1. Ice isn't melting
1.2. Heading into ice age
1.3. Weather is cold
1.4. Hiatus in warming
1.5. Oceans are cooling
1.6. Sea level rise is exaggerated
1.7. Extremes aren't increasing
1.8. Changed the name
2. Human Greenhouse Gases are not causing global warming
2.1. It's natural cycles
2.2. Non-Greenhouse Gas forcings
2.3. No evidence for Greenhouse Effect
2.4. CO_2 not rising
2.5. Emissions not raising CO_2 levels
Level 1 claims
Level 2 claims
3. Climate impacts are not bad
3.1. Sensitivity is low
3.2. No species impact
3.3. Not a pollutant
3.4. Only a few degrees
3.5. No link to conflict
3.6. No health impacts
4. Climate solutions won't work
4.1. Policies are harmful
4.2. Policies are ineffective
4.3. Too hard
4.4. Clean energy won't work
4.5. We need energy
5. Climate movement/science is unreliable
5.1. Science is unreliable
5.2. Movement is unreliable
5.3. Climate is conspiracy

Coan et al. (2021)

Figure 2: CARDS taxonomy of contrarian climate claims (Coan et al., 2021).

puter Assisted Recognition of Denial and Skepticism)—was trained to detect specific contrarian claims about climate change (Coan et al., 2021). To achieve this, a taxonomy of contrarian claims about climate change was developed, described as the CARDS taxonomy (see Figure 2). In contrast to the technique-based FLICC taxonomy, the CARDS taxonomy takes a fact-based approach, examining the content claims in contrarian arguments. The CARDS model has been found to be successful in detecting specific content claims in contrarian blogs and conservative think-tank articles (Coan et al., 2021) as well as in climate tweets (Rojas et al., 2023).

While the CARDS model was developed in order to facilitate automatic debunking of climate misinformation, it by design was only able to detect content-claims. Flack et al. (2023) found that contrarian claims in the CARDS taxonomy often contain multiple logical fallacies. As an effective debunking needs to contain both explanation of the facts and the misinformation's fallacies (Lewandowsky et al., 2020), automated detection of climate misinformation needs to include not only content-claim detection such as that provided by the CARDS model but also detect any fallacies contained in the misinformation.

There have been several studies that have used machine learning to attempt detecting logical fallacies in climate-themed text. Jin et al. (2022) developed a structure-aware model to detect fallacies in both climate text and general text, arguing that the task is about the "form" or "structure" of the argument rather than the content words. However, as indicated in Table 1, certain fallacies do not rigidly adhere to a fixed structure. Instead, their detection necessitates a background knowledge base. Alternatively, Alhindi et al. (2023) used instruction-based prompting to detect 28 fallacies across a range of topics, including climate change. These past efforts have shown low accuracy in fallacy detection and the fallacy frameworks used showed little overlap with the FLICC and CARDS frameworks that have been developed specifically to facilitate detection and debunking of climate misinformation. Moreover, upon meticulous examination of Jin et al. (2022) and Alhindi et al. (2023). datasets, available on [1] and [2] respectively, several data quality issues were identified. These included the presence of duplicate samples, instances of duplicate samples bearing different labels, repetition of samples across the training, validation, and test sets, label merging, empty samples, and ultimately, discrepancies between our formulated fallacy definitions and their annotations.

This study integrates past psychological, critical thinking, and computer science research in order to develop a technocognitive solution to fallacy detection. Technocognition is the synthesis of psychological and technological research in order to develop holistic, interdisciplinary solutions to misinformation

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

(Lewandowsky et al., 2017). By synthesising the CARDS and FLICC framework, we will develop an interdisciplinary solution to fallacy detection that can then be implemented in a generative debunking solution, bringing this research closer to the “holy grail of fact-checking”.

2 Methods

2.1 Developing a FLICC/CARDS dataset

We developed a training dataset that mapped examples of climate misinformation to fallacies from the FLICC taxonomy as well as the contrarian claim in the CARDS taxonomy. Text was manually taken from several datasets - the contrarian blogs and CTT articles in the Coan et al. (2021) training set, the climate datasets from Alhindi et al. (2023) and Jin et al. (2022), and the test set of climate tweets from Rojas et al. (2023). In order to more reliably identify dominant fallacies in text, the critical thinking methodology from Cook et al. (2018) was used to deconstruct difficult examples. A selection of sample deconstructions of the most common combinations of CARDS claims and FLICC fallacies are shown in Table 2.

To further ensure the quality of our manually annotated dataset, we conducted a rigorous examination of our samples. First, we searched for potential duplicates by employing exact matching techniques. Subsequently, we leveraged Bert embeddings (Devlin et al., 2019) to construct a similarity matrix, utilising cosine similarity (equation 1) as the measure of similarity between samples. We then manually reviewed both the exact matches and pairs of samples with the highest similarity scores and proceeded to remove them. For instance, we identified identical and seemingly identical samples that differed only in extra whitespaces, punctuation marks, or capitalization. We also encountered similar texts referring to distinct records, places, or dates, and in such cases, we retained the most representative of these samples.

$$\cos\varphi = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

$$d\left(p,q\right) = \sqrt{p\cdot p - 2(p\cdot q) + q\cdot q} \tag{2}$$

In addition to identifying duplicate samples, we aimed to detect outliers, recognising the possibility of inadvertent misannotation of sample labels. Utilising the same Bert embeddings from before, we calculated the mean embedding for each unique label category. Next, we calculated the euclidean distance (equation 2) of all samples associated with a particular label from its corresponding mean embedding. We selected 36 samples with notably larger distances. Furthermore, we applied the Isolation Forest algorithm (Liu et al., 2008), a robust technique for outlier detection, and identified a set of 50 potential outliers which included the 36 samples identified earlier. Out of these 50 outliers, we didn’t find misannotated labels, but we selectively removed four samples, primarily for being confusingly worded.

The dataset offered a deeper insight into the interplay between FLICC fallacies and CARDS claims, shown in Figure 3. It showed a much broader distribution of fallacies within each CARDS claim than found in Flack et al. (2023). This indicated that contrarian arguments could take various forms featuring different fallacies, and that merely detecting a CARDS claim was not sufficient in identifying the argument’s fallacy. This underscored the imperative of developing a model for reliably detecting FLICC fallacies in climate misinformation. Our process resulted in a dataset of 2509 samples.

2.2 Training a Model to Detect Fallacies

2.2.1 Model selection

Classifying fallacies, especially when they revolve around a singular subject such as climate change, poses a significant challenge. Jin et al. (2022) contended that this classification task primarily concerned the “form” or “structure” of the argument rather than the specific content words used. Yet, as depicted in Figure 3, it becomes evident that certain fallacies exhibit a higher prevalence within specific claims.

From the array of available tools, we hypothesised that the low-rank adaptation (LoRa) approach (Hu et al., 2021) might offer a promising initial solution to our problem. LoRa brings several advantages

Misinformation Example	Claim	Deconstruction	Fallacy Explanation
"In many environmental fields, the science is being abused and distorted to promote a political and financial agenda"	5.2	P1: Environmental science is being abused and distorted. P2: The distortion of science is to promote political/financial agendas. C: Environmental science is discredited.	P1 commits ad hominem, accusing science of deceptive or unethical acts. P2 commits misrepresentation as it assumes science is driven by a political agenda.
"Yet another global warming alert, when global temperatures are heading down and records for cold are being broken left, right and center."	1.3	P1: Cold weather events are occurring. HP: If global warming was happening, we wouldn't experience cold events. C: Global warming isn't happening.	HP commits anecdote, using isolated incidents limited in place and time to make conclusions about global warming. HP also commits impossible expectations as cold events will continue to happen under global warming but they are less likely to happen while hot events are more likely to occur.
"The most extraordinary fraud in the history of Western science: the fantasy that by controlling anthropogenic emissions of carbon dioxide, mankind can control global temperatures."	5.3	P1: Scientists have commited a range of conspiratorial actions to defend the mainstream view and suppress dissenting views. C: There is a conspiracy among scientists to deceive the public.	P1 commits conspiracy theory, assuming that there is secret plotting behind climate science and that scientists act with nefarious intent.
"Yes, there is climate change happening. The world's climate always changes."	2.1	P1: Climate has changed due to natural causes in the Earth's past. P2: Climate is changing now. HP: What caused climate change in the past must be the same as what's causing climate change now. C: Current climate change must be natural.	HP commits single cause, assuming that what caused climate change in the past (natural factors) must be the same as what's causing climate change now.
"We, the animals and all land plant life would be healthier if CO2 content were to increase."	3.3	P1: CO2 is beneficial for plant growth. HP: Increased CO2 only has beneficial effects for plants. C: Emitting more CO2 will be good for plants.	HP commits oversimplification, ignoring the ways that climate change impacts agriculture through increased heat stress and flooding. CO2 fertilisation is just one factor affecting plant growth. The full picture shows that negative impacts outweigh benefits.
"CO2 is incapable of causing climatic warming by itself. CO2 makes up only 0.038% of the atmosphere and accounts for only a few percent of the greenhouse gas effect."	2.3	P1: CO2 is a trace gas, comprising only a small component of the atmosphere. HP: If there is a small percentage of CO2 in the atmosphere, its warming potential is low. C: CO2 cannot be the main cause of global warming.	HP commits misrepresentation as small active substances can have a strong effect (e.g., it only takes a small amount of mercury to poison someone).
"Sea ice is setting records this year."	1.1	P1: In the short term, Arctic sea ice hasn't changed much. HP: If Arctic sea ice maximum extent hasn't changed much in the short term, then Arctic sea ice is fine in the long-term. C: Arctic sea ice is fine.	HP commits cherry picking, looking at a short period of sea ice data while ignoring the long-term decline in Arctic sea ice.

Table 2: Deconstructions of climate misinformation examples (seven most common FLICC/CARDS combinations/.

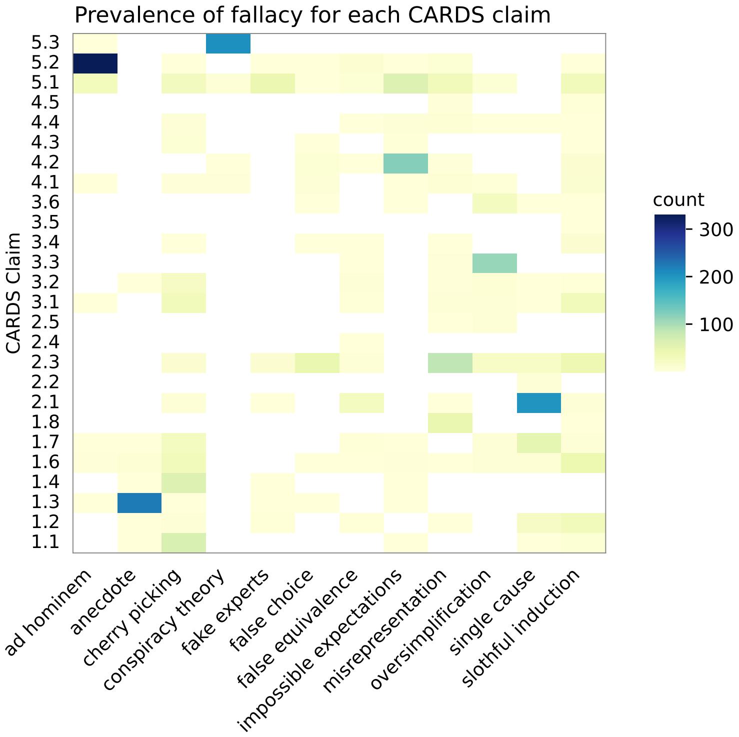


Figure 3: Map of fallacies across different CARDS claims.

in terms of storage and hardware efficiency when adapting large language models to downstream tasks. What captivated our interest was how adapting the model weights through trainable rank decomposition matrices could be beneficial for our segmentation problem.

In order to test our hypothesis, we evaluated all accessible models within HuggingFace's Parameter-Efficient Fine-Tuning (PEFT) library [3] for sequence classification, with the exclusion of GPT-J due to hardware limitations. Specifically, we tested the following model checkpoints: *bert-base-uncased*,*roberta-large*, *gpt2*, *bigscience/bloom-560m*, *facebook/opt-350m*, *EleutherAI/gpt-neo-1.3B*, *microsoft/deberta-base*, *microsoft/deberta-v2-xlarge*.

2.2.2 Experimental setup

We employed the PyTorch [4] framework and HuggingFace [5] libraries for our experiments, conducting an iterative analysis to determine the optimal configuration at each experimental stage. Our dataset was partitioned into train, validation, and test sets as illustrated in Table 3. The models were trained for a maximum of 30 epochs, and we utilised the validation set to mitigate overfitting by employing an early stopping method after three consecutive rounds without improvement. For each experiment, out of all the training epochs, we selected the model with the best F1-macro score, considering the imbalanced nature of our dataset.

[3] https://github.com/huggingface/peft

[4] https://pytorch.org

[5] https://huggingface.co

Label	train	val	test	Total
ad hominem	264	67	37	368
anecdote	170	43	24	237
cherry picking	222	56	31	309
conspiracy theory	154	39	22	215
fake experts	44	12	7	63
false choice	48	13	7	68
false equivalence	52	14	8	74
impossible expectations	144	37	21	202
misrepresentation	151	38	22	211
oversimplification	143	36	20	199
single cause	226	57	32	315
slothful induction	178	45	25	248
Total	1,796	457	256	2,509

Table 3: Fallacy types and their number of samples on each partition in the FLICC dataset.

We examined the best learning rates within 1.0e-5, 5.0e-5 and 1.0e-4. We fixed the batch size at 32, employed the AdamW optimiser with a weight decay of 0.0, and utilised the cross-entropy loss function. Once we determined the best learning rate for the model, we moved to the second round of experiments using focal loss (Lin et al., 2018) instead of cross-entropy loss. Focal loss enables the emphasis on harder-to-classify samples by introducing a gamma penalty to the results; we analysed gamma values of 2, 4, 6, and 16.

Subsequently, we completed a third round of experiments by adding the weight decay parameter, exploring values of 0.1 and 0.01. Again, we did it for the best model identified previously, either with or without focal loss. Finally, we conducted a fourth round of experiments testing LoRa ranks of 8 and 16, as well as alpha values of 8 and 16.

3 Results

3.1 Baseline

The initial step involved establishing a ZeroR classifier, i.e., a classifier that always selects the most frequent class. In our test set, which comprises 256 samples, the most frequent label is “Ad Hominem” with 37 samples. The ZeroR classifier achieved an accuracy of 0.14 and macro f1 score of 0.02.

3.1.1 Comparing our model to Google’s Palm2 / OpenAi’s GPT3.5

General-purpose LLMs available online exhibit limited proficiency in specific tasks such as fallacy detection. We conducted an evaluation by applying our test set of 256 test samples to Google’s Palm2 and OpenAI’s GPT-3.5 via their respective APIs. We used the following prompt: “Please classify a piece of text into the following categories of logical fallacies: [a list of all logical fallacy types]. Text: [Input text] Label: ”

The attained overall accuracy scores for Palm2 and GPT-3.5 in detecting labels were 0.25 and 0.21, respectively. In a detailed analysis of these results, Palm2 failed to assign a label to 28 out of 256 samples, while GPT-3.5 left seven samples without predictions. In both cases, the models produced responses such as “This text does not contain any logical fallacies” or “None of the above.” The most common predictions for both models were “false equivalence” and “cherry picking.” GPT-3.5 exhibited a pronounced bias towards predicting “false equivalence” (102) and “cherry picking” (49), whereas Palm2 leaned towards “false equivalence” (44) and “cherry picking” (55). The detailed breakdown of the complete prediction results can be found in Table 4.

	Palm2			Gpt-3.5		
	P	R	F_1	P	R	F_1
ad hominem	0.00	0.00	0.00	0.77	0.27	0.40
anecdote	0.50	0.04	0.08	0.57	0.17	0.26
cherry picking	0.38	0.68	0.49	0.22	0.35	0.27
conspiracy theory	0.64	0.41	0.50	0.60	0.55	0.57
fake experts	0.62	0.71	0.67	0.40	0.29	0.33
false choice	0.14	0.43	0.21	0.40	0.29	0.33
false equivalence	0.05	0.25	0.08	0.04	0.50	0.07
impossible expectations	0.40	0.10	0.15	0.25	0.24	0.24
misrepresentation	0.26	0.27	0.27	0.00	0.00	0.00
oversimplification	0.25	0.05	0.08	0.12	0.15	0.13
single cause	0.40	0.31	0.35	1.00	0.03	0.06
slothful induction	0.15	0.16	0.16	0.00	0.00	0.00
accuracy			0.25			0.21
macro avg	0.29	0.26	0.23	0.34	0.22	0.21
weighted avg	0.31	0.25	0.24	0.42	0.21	0.22

Table 4: Classification results for Palm2-text-bison-001 (Palm2) and gpt-3.5-turbo-instruct (Gpt-3.5). For each class, we report precision (P), recall (R), and F_1 score.

3.2 Assessing our model performance at detecting different fallacies

Table 5 summarises test f1-macro score results for all the analysed models. The poor performance of the LoRa experiments was surprising. Only *roberta-large* and *bigscience/bloom-560m* succeeded in attaining f1-macro scores comparable to those from previous settings. However, neither of these experiments outperformed the previously achieved scores, indicating possible areas for future work.

	Learning rate			Focal loss, gamma param.				Weight decay		LoRa	
Model checkpoints	1.0E-05	5.0E-05	1.0E-04	2	4	8	12	0.01	0.10	8	16
bert-base-uncased	0.56	0.65	0.58	0.64	0.61	0.63	0.56	0.64	0.62	0.36	0.37
roberta-large	0.66	0.68	0.02	0.01	0.00	0.69	0.00	0.01	0.00	0.60	0.64
gpt2	0.42	0.56	0.47	0.51	0.45	0.46	0.46	0.57	0.50	0.10	0.30
bigscience/bloom-560m	0.54	0.54	0.33	0.48	0.50	0.56	0.52	0.46	0.51	0.44	0.44
facebook/opt-350m	0.23	0.12	0.02	0.20	0.23	0.22	0.22	0.21	0.22	0.07	0.07
EleutherAI/gpt-neo-1.3B	0.44	0.65	0.58	0.44	0.05	0.50	0.49	0.57	0.57	0.33	0.33
microsoft/deberta-base	0.67	0.63	0.62	0.64	0.63	0.65	0.56	0.69	0.67	0.02	0.02
microsoft/deberta-base-v2-xlarge	0.67	0.41	0.02	0.70	0.73	0.63	0.69	**0.73**	0.71	0.07	0.38

Table 5: F1 macro scores, highlighted in green are the best model parameter combination for each model. Best model overall is microsoft/deberta-base-v2-xlarge, lr=1.0e-5, gamma=4, weight decay=0.01 fine-tuned over 15 epochs.

The most effective model overall was microsoft/deberta-base-v2-xlarge (He et al., 2021) with a learning rate of 1.0e-5, focal loss with gamma penalty of 4, weight decay of 0.01, and fine-tuned by 15 epochs. The detailed breakdown of the results can be found in Table 6, with the small gap between validation and test results indicating the model's ability to generalise effectively. Table 7 displays the confusion matrix, depicting actual labels on the y-axis and predicted labels on the x-axis. Greater F1-score performance was observed for fake experts, anecdote, conspiracy theory and ad hominem. In contrast, false equivalence and slothful induction exhibited the lowest F1-scores.

	Validation			Test		
	P	R	F_1	P	R	F_1
ad hominem	0.76	0.75	0.75	0.81	0.78	0.79
anecdote	0.95	0.86	0.90	0.88	0.92	0.90
cherry picking	0.69	0.66	0.67	0.77	0.77	0.77
conspiracy theory	0.78	0.82	0.80	0.78	0.82	0.80
fake experts	1.00	0.92	0.96	1.00	1.00	1.00
false choice	0.83	0.77	0.80	0.62	0.71	0.67
false equivalence	0.50	0.43	0.46	0.50	0.38	0.43
impossible expectations	0.69	0.73	0.71	0.69	0.86	0.77
misrepresentation	0.63	0.63	0.63	0.68	0.68	0.68
oversimplification	0.88	0.58	0.70	0.78	0.70	0.74
single cause	0.81	0.74	0.77	0.81	0.66	0.72
slothful induction	0.54	0.82	0.65	0.50	0.56	0.53
accuracy			0.73			0.74
macro avg	0.75	0.73	0.73	0.74	0.74	0.73
weighted avg	0.75	0.73	0.73	0.75	0.74	0.74

Table 6: Classification report. For each class, we report precision (P), recall (R), F_1 score for validation and test partitions.

ad hominem	0.78		0.03	0.11				0.03	0.03			0.03
anecdote		0.92									0.04	0.04
cherry picking	0.03		0.77			0.03			0.03	0.03	0.03	0.06
conspiracy theory	0.14			0.82			0.05					
fake experts					1.00							
false choice	0.14					0.71						0.14
false equivalence	0.13						0.38	0.25			0.25	
impossible expectations								0.86	0.10			0.05
misrepresentation				0.05				0.14	0.68	0.09		0.05
oversimplification			0.05						0.05	0.70		0.20
single cause		0.09	0.06				0.06	0.03			0.66	0.09
slothful induction	0.04		0.12			0.08		0.04	0.08	0.04	0.04	0.56

Table 7: Normalised confusion matrix (actual labels on y-axis, predicted labels on x-axis)

3.2.1 Comparing our model to Alhindi and Jin

Although the comparison is not straightforward, both Jin et al. (2022) and Alhindi et al. (2023) developed datasets of climate change fallacies and trained machine learning models with similar numbers of fallacies (13 and 9 respectively). They reported overall F1-scores of 0.21 and 0.29 for their climate datasets in their best round of experiments, whereas we achieved an F1-score 0.73. Direct comparison between these studies and our results are difficult as we do not share the same set of fallacies, but Table 8 provides a summary of the results for the shared fallacies between the scores obtained by Jin et al. (2022) and Alhindi et al. (2023) using their respective models on their datasets, and our model's performance on our dataset.

Alhindi et al. (2023)	max. F_1	F_1	FLICC
causal oversimplification	0.53	**0.72**	single cause
cherry picking	0.43	**0.77**	cherry Picking
irrelevant authority	0.30	**1.00**	fake experts

Jin et al. (2022)	F_1	F_1	FLICC
intentional	0.25	**0.77**	cherry picking
ad hominem	0.42	**0.79**	ad hominem
false dilemma	0.17	**0.67**	false choice

Table 8: Summary of comparable labels (fallacies)

4 Discussion

In this study, we developed a model for classifying logical fallacies in climate misinformation. Our model showed good performance in classifying a dozen fallacies. The results showed significant improvement on previous efforts to detect climate fallacies. The Deberta model also showed better results than those found with Palm2 and GPT3.5 models. An interactive tool has been made available online allowing users to enter text and receive model predictions at https://huggingface.co/fzanartu/flicc

Nevertheless, our model showed lower performance with some fallacies relative to others. The false equivalence fallacy showed lowest performance, which can likely be explained by the relative lack of training examples. However, this factor cannot explain the low performance of slothful induction, which had a relatively high number of training examples. One contributor to the difficulty in detecting slothful induction was the conceptual overlap between slothful induction and cherry picking. Both fallacies involve coming to conclusions by ignoring relevant evidence when coming to a conclusion but cherry picking achieves this through an act of commission (citing a narrow piece of evidence that conflicts with the full body of evidence) while slothful induction uses an act of omission (coming to conclusions without citing evidence). Another factor to consider in analysing the poor performance of slothful induction as illustrated in Figure 3 is that the labels of slothful induction and cherry picking stand out as the most widely represented across various topics in CARDS claims. However, cherry picking is concentrated in fewer claims compared to slothful induction, which exhibits a more even distribution across all claim topics.

Another source of difficulty are texts that contain multiple fallacies. It's common that climate misinformation incorporates several elements in a single item. An example is making a content claim such as "a cooling sun will stop global warming" while also including an ad hominem attack against "alarmists". Other research also struggled with the fact that climate misinformation often contains multiple claims, necessitating the need for multi-label classification (Coan et al., 2021). Further, some texts may include a single claim that nevertheless contains multiple fallacies. For example, the claim that "there's no evidence that CO2 drove temperature over the last 400,000 years" commits slothful induction by ignoring all the evidence for CO2 warming as well as false choice by demanding that either CO2 drives temperature or temperature drives CO2 (Flack et al., 2023).

Future research could look to improve the model's performance by increasing the number of training examples, particularly for underrepresented fallacies such as false equivalence, fake experts, and false choice. As an active area of research, exploring additional or novel classification models and methodologies, such as LoRa, remains an option. However, our primary interest lies in developing a more comprehensive approach that could potentially bring us closer to the "holy grail of fact-checking" a more adept understanding of our deconstructive methodology and imitation of critical thinking within large language models (LLMs). One potentially more accessible avenue involves creating an automated ReAct agent (Yao et al., 2023) that we can further optimise using evolutionary computation techniques, as detailed in (Fernando et al., 2023). A more sustainable, long-term approach might involve fine-tuning a LLM, following the methodologies and findings outlined in An et al. (2023) and Huang et al. (2023).

This study restricted its scope to climate misinformation and fallacies used within contrarian claims about climate change. However, the FLICC taxonomy has also been applied to other topics such as vaccine misinformation (Hopkins et al., 2023). The model could be generalised to tackle general misinformation or other specific topics.

Future research could explore combining our fallacy detection model with models that detect contrarian CARDS claims (Coan et al., 2021; Rojas et al., 2023). Potentially, a model that can detect both content claims in climate misinformation and fallacies could generate debunkings that adhere to the fact-myth-fallacy structure recommended by psychological research (Lewandowsky et al., 2020).

The issues the model faced with texts that contain multiple fallacies point to an important area of interaction between computer and cognitive science. When misinformation contain multiple fallacies, what is the ideal response from a communication approach? Past analysis has found that climate misinformation frequently contains multiple fallacies (Cook et al., 2018; Flack et al., 2023). While there is indication that corrections that explain two fallacies are more effective than single-fallacy corrections (Hayes et al., 2023), there is a dearth of other research exploring the optimal communication approach for countering misinformation with multiple fallacies. Figure 3 illustrates that contrarian climate claims can commit a number of fallacies and as technology to detect these fallacies improves, communication science will need to progress to inform on optimal response strategies.

This interaction between psychological and computer science research illustrates the value of the technocognitive approach to misinformation research. Inevitably, many technological solutions will eventually need to interact with humans, at which time psychological factors need to be understood to ensure the interventions are effective. Our model was built from frameworks developed in a body of psychological and critical thinking work (Coan et al., 2021; Cook et al., 2018, 2017; Vraga et al., 2020), and the output from the model will eventually be implemented in communication informed by psychological research.

References

Alhindi, T., Chakrabarty, T., Musi, E., and Muresan, S. (2023). Multitask instruction-based prompting for fallacy recognition.

An, S., Ma, Z., Lin, Z., Zheng, N., Lou, J.-G., and Chen, W. (2023). Learning from mistakes makes llm better reasoner.

Banas, J. A. and Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, 39(2):184–207.

Boussalis, C. and Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change*, 36:89–100.

Coan, T. G., Boussalis, C., Cook, J., and Nanko, M. O. (2021). Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Cook, J. (2020). Deconstructing climate science denial. *Research Handbook on Communicating Climate Change*, pages 62–78.

Cook, J., Ellerton, P., and Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, 13(2):024018.

Cook, J., Lewandowsky, S., and Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one*, 12(5):e0175799.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Diethelm, P. and McKee, M. (2009). Denialism: what is it and how should scientists respond? *The European Journal of Public Health*, 19(1):2–4.

Ecker, U. K., Lewandowsky, S., and Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, 38:1087–1100.

Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences*, 113(1):92–97.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rocktäschel, T. (2023). Promptbreeder: Self-referential self-improvement via prompt evolution.

Flack, R., Cook, J., Ellerton, P., Kinkead, D., Coan, T., Boussalis, C., Nanko, M., Gallant, A., and Dargaville, R. (2023). Identifying reasoning fallacies in a comprehensive taxonomy of contrarian claims about climate change. *Environmental Communications*.

Geiger, N. and Swim, J. K. (2016). Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *Journal of Environmental Psychology*, 47:79–90.

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., and Yu, C. (2015). The quest to automate fact-checking. In *Proceedings of the 2015 computation+ journalism symposium*. Citeseer.

Hayes, O. R., Lieu, R., and Cook, J. (2023). Testing the impact of fallacies and technique-based corrections in climate change misinformation.

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Hopkins, K. L., Lepage, C., Cook, W. M., Thomson, A., Abeyesekera, S., Knobler, S., Boehman, N., Thompson, B., Waiswa, P., Ssanyu, J. N., Kabwijamu, L., Wamalwa, B., Aura, C., Rukundo, J. C., and Cook, J. (2023). Co-designing a mobile-based game to improve misinformation resistance and vaccine knowledge in uganda, kenya, and rwanda. *Journal of Health Communication*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2023). Large language models cannot self-correct reasoning yet.

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., and Schölkopf, B. (2022). Logical fallacy detection.

Lewandowsky, S., Cook, J., Ecker, U., Albarracín, D., Kendeou, P., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., et al. (2020). The debunking handbook 2020.

Lewandowsky, S., Ecker, U. K., and Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). Focal loss for dense object detection.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

McCright, A. M., Charters, M., Dentzman, K., and Dietz, T. (2016). Examining the effectiveness of climate change frames in the face of a climate change denial counter-frame. *Topics in cognitive science*, 8(1):76–97.

Ranney, M. A. and Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in cognitive science*, 8(1):49–75.

Rojas, C., Algra-Maschio, F., Andrejevic, M., Coan, T., Cook, J., and Li, Y.-F. (2023). Augmented cards: A machine learning approach to identifying triggers of climate change misinformation on twitter.

Schmid, P. and Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, 3(9):931–939.

Stecula, D. A. and Merkley, E. (2019). Framing climate change: Economics, ideology, and uncertainty in american news media content from 1988 to 2014. *Frontiers in Communication*, 4:6.

Van der Linden, S., Leiserowitz, A., Rosenthal, S., and Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global challenges*, 1(2):1600008.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *science*, 359(6380):1146–1151.

Vraga, E. K., Kim, S. C., Cook, J., and Bode, L. (2020). Testing the effectiveness of correction placement and type on instagram. *The International Journal of Press/Politics*, 25(4):632–652.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models.