www.nature.com/scientificreports

scientific reports

OPEN

A technocognitive approach to detecting fallacies in climate misinformation

Francisco Zanartu[1], John Cook[2]✉, Markus Wagner[3] & Julian García[3]

Misinformation about climate change is a complex societal issue that requires holistic, 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 that counter misinformation require both fact-based (e.g., factual explanations) and technique-based (e.g., explanations of misleading techniques and logical fallacies) 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 examples of climate misinformation to reasoning fallacies. This dataset is used to train a model to detect fallacies in climate misinformation. We evaluate the model's performance using the F_1 score, which measures how well the model detects relevant cases while avoiding irrelevant ones. Our study shows F_1 scores that are 2.5–3.5 times better than previous works. The fallacies that are easiest to detect include fake experts and anecdotal arguments, while 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.

Keywords Fallacy detection, Climate change, Technocognitive approach, Misinformation

Misinformation about climate change reduces climate literacy and undermines support for policies that mitigate climate impacts[1] while exacerbating public polarization[2]. Efforts to communicate the reality of climate change can be canceled out by misinformation[3]. Ignorance about the strong degree of public acceptance about the reality of climate change is associated with "climate silence"[4]. These impacts necessitate interventions that neutralize their negative influence.

A growing body of psychological research has tested a variety of interventions aimed at reducing the impact of misinformation[5]. Two leading communication approaches are fact-based and technique-based. Fact-based corrections—also described as topic-based[6]—involve exposing how misinformation is false through factual explanations. Technique-based corrections—also described as logic-based[7,8]—involve explaining misleading rhetorical techniques and logical fallacies used in misinformation. One study found that both fact-based and technique-based corrections were effective in countering misinformation[6]. However, technique-based corrections have also been found to outperform fact-based corrections as they were equally effective whether the correction was encountered before or after the misinformation, while fact-based corrections were ineffective if misinformation was shown afterwards, leading to a canceling out effect[8]. This result is consistent with other studies finding that factual explanations can be cancelled out if encountered alongside contradicting misinformation[2,3,9]. Technique-based interventions can also address misinformation techniques such as paltering or cherry picking which use factual statements to mislead by withholding relevant information[10]. By 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[11].

Consequently, increasing research attention has focused 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[12]—summarised with the acronym

[1]University of Melbourne, Parkville, VIC, Australia. [2]Melbourne Centre for Behaviour Change, University of Melbourne, Parkville, VIC, Australia. [3]Department of Data Science & AI, Monash University, Clayton, VIC 3800, Australia. ✉email: jocook@unimelb.edu.au

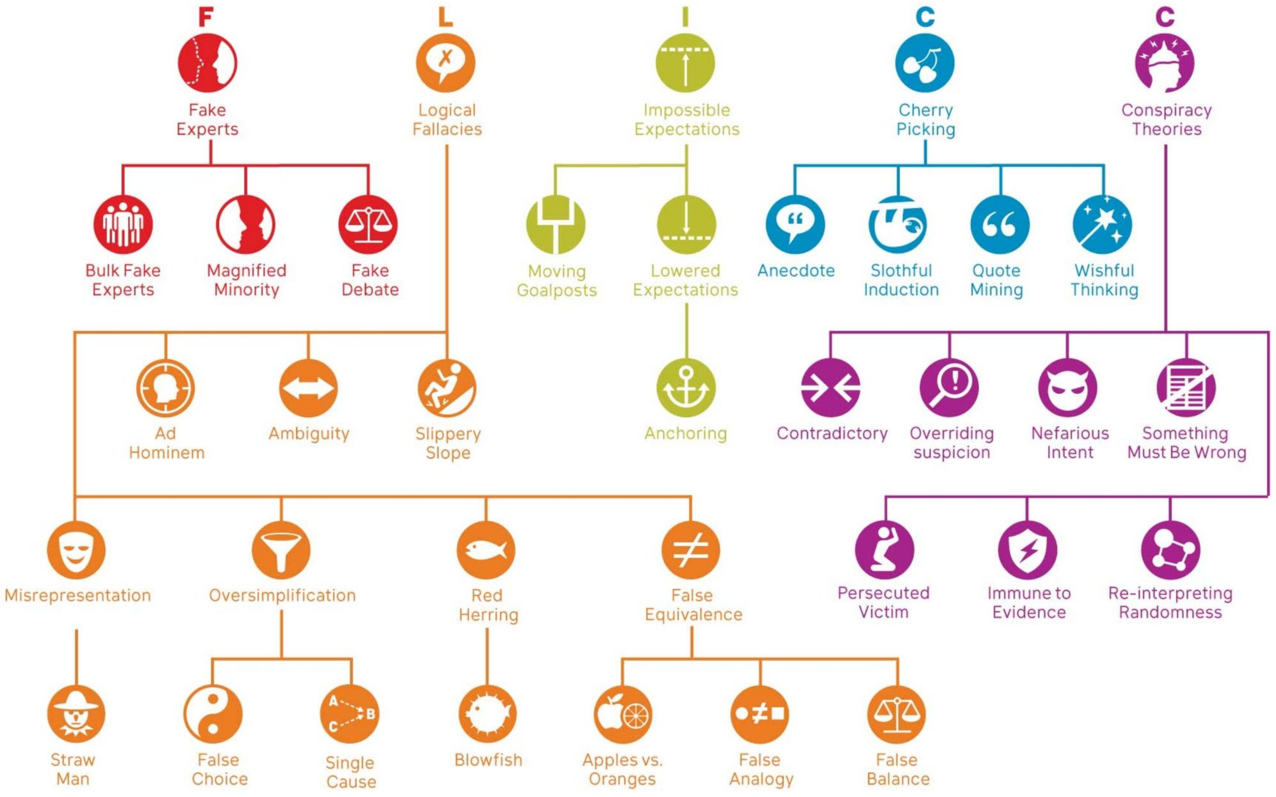


Fig. 1. FLICC taxonomy of misinformation techniques and logical fallacies[13].

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.	P has expertise in a non-climate topic. Therefore, P is an expert on climate
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.

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 Fig. 1[13]. A critical thinking methodology was developed for manually deconstructing and analysing climate misinformation in order to identify misleading logical fallacies[14]. This methodology has been applied to contrarian climate claims in order to identify the fallacies used in specific climate myths[15]. Table 1 lists the fallacies identified in climate misinformation, as well as their definitions. 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 logical structure of each fallacious argument.

While these theoretical frameworks have been developed based on psychological and critical thinking research, developing practical solutions countering misinformation is challenging for various reasons. The public perceives misinformation as more novel than factual information, resulting in it spreading faster and farther through social networks than true news[16]. Further, people continue to be influenced by misinformation, even if they remember a retraction-a phenomenon known as the continued influence effect[17]. To address these challenges, research has begun to focus on pre-emptive or rapid response solutions such as inoculation or misconception-based learning[18].

One proposed solution is automatic and instantaneous detection and fact-checking of misinformation, described as the "holy grail of fact-checking"[19]. Machine learning models offer a tool towards achieving this goal. For example, 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[20]. Similarly, topic modelling has been combined with network analysis to find an association between corporate funding and polarizing climate text[21]. Lastly, topic modelling of newspaper articles has been used to identify economic or uncertainty framing about climate change[22]. While the unsupervised approach offers general insights about the nature of climate misinformation with large datasets, it does not facilitate detection of specific misinformation claims which is necessary in order to generate automated fact-checks.

To address this shortcoming, a supervised machine model—the CARDS model (Computer Assisted Recognition of Denial and Skepticism)—was trained to detect specific contrarian claims about climate change[23]. To achieve this, the CARDS taxonomy was developed, organizing contrarian claims about climate change into hierarchical categories (see Fig. 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[23] as well as in climate tweets[24].

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.[15] found that contrarian claims in the CARDS taxonomy often contained multiple logical fallacies. As an effective debunking needs to contain both explanation of the facts and the fallacies employed by the misinformation[11], 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.

Several studies have utilized machine learning to detect logical fallacies in climate-themed text.[25] developed a structure-aware model to detect fallacies in both climate text and general text, emphasising the importance of the argument's form or structure over its content words. However, certain fallacies, as indicated in Table 1, do not strictly adhere to a fixed structure, requiring a background knowledge base for detection. Alternatively,[26]

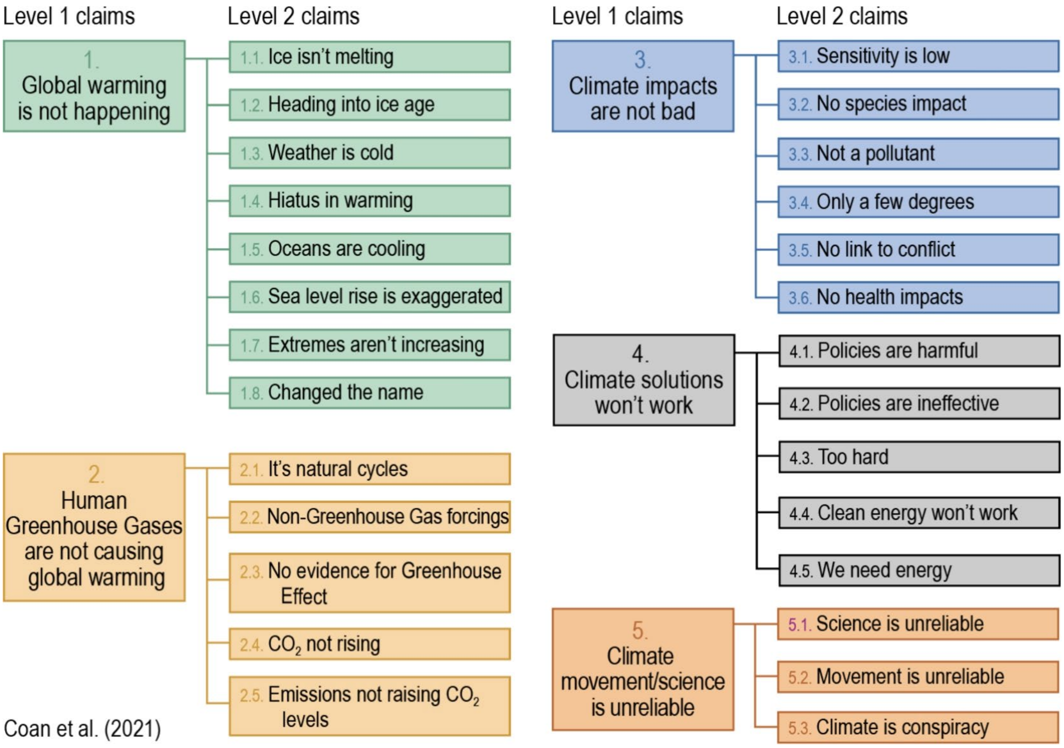

Level 1 claims
Level 2 claims
1. Global warming is not happening
1.1. Ice isn't melting
1.2. Heading into ice age
1.3. Weather is cold
1.4. Hiatus in warming
1.5. Oceans are cooling
1.6. Sea level rise is exaggerated
1.7. Extremes aren't increasing
1.8. Changed the name
2. Human Greenhouse Gases are 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. CO₂ not rising
2.5. Emissions not raising CO₂ levels
Coan et al. (2021)
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

Fig. 2. CARDS taxonomy of contrarian climate claims[23].

employed instruction-based prompting to detect 28 fallacies across a range of topics, including climate change. Despite these efforts, past studies have demonstrated low accuracy in fallacy detection, and the frameworks used showed limited overlap with FLICC and CARDS frameworks specifically developed for climate misinformation detection and debunking. After closely examining the datasets from[25] and[26], which are available at (https://github.com/causalNLP/logical-fallacy) and (https://github.com/Tariq60/fallacy-detection), we found several data quality issues. These issues included duplicate samples, instances of duplicate samples with different labels, sample repetition across training, validation, and test sets, label merging, empty samples, and ultimately, discrepancies between our formulated fallacy definitions and their annotations.

Our study integrated 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[27]. For example, digital games such as Bad News[28] and Cranky Uncle[29] apply inoculation theory in interactive games that build public resilience against misinformation. By synthesising the CARDS and FLICC framework, we developed an interdisciplinary solution to fallacy detection that could subsequently be implemented in automated debunking solutions, bringing this research closer to the "holy grail of fact-checking".

Results

Baseline

The initial step involved establishing a ZeroR classifier, i.e., a classifier that always selects the most frequent class. Our test set comprised a stratified random sampling, where the most frequent label is "Ad Hominem", occurring 37 times out of 256 instances. We present the derived accuracy of 0.14 and F_1 scores of 0.02. These scores can be calculated by employing the respective formula 1 for the accuracy score and 2 for the F_1 score where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives, and FP is the number of false positives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{2}$$

Comparing our model to Google's Gemini and OpenAi's GPT-4

Assessing the reasoning skills of large language models (LLMs) is an active area of research, where natural language inference is one of their hardest tasks. One of our goals was to compare our tool to LLMs by applying our test set of 256 samples to Google's Gemini (Gemini-1.0-pro)[30] and OpenAI's GPT-4 (GPT-4-0125-preview)[31] using 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 overall accuracy scores for Gemini-pro and GPT-4 in detecting labels were 0.21 and 0.32, both surpassing the ZeroR classifier by 1.5 and 2.3 times. Although LLMs showed an improvement over the most simple baseline, still far from being a reliable tool for this task. In a detailed analysis of these results, Gemini-pro failed to label eight out of the 256 samples with empty responses or replying "None of the above". Gemini-pro's most common predictions were "Oversimplification" (158), "Conspiracy theory" (45) and "Cherry picking" (20). Also, the safety settings were disabled in order to obtain Gemini-pro predictions, as some myths were blocked by the API.

GPT-4, on the other hand, failed to label 44 out of the 256 samples by providing unrequested information and comments such as "... the closest interpretation could be cherry picking" or "The provided text does not seem to fall into any of the listed categories ... Label: None". In these cases, the most likely label was assigned so that in the examples above, the label would be "cherry picking" and "None." With that consideration, GPT-4 assigned "None" to four samples. Its most frequent predictions were "Oversimplification" (84), "Conspiracy theory" (38) and "Anecdote" (26). Table 2 shows the detailed break down of results.

Assessing our model performance at detecting different fallacies

Table 3 summarises test F_1-macro score results for all the analysed models. The poor performance of the Low-Rank Adaptation(LoRa)[32] experiments was surprising. Only *roberta-large* and *bigscience/bloom-560m* succeeded in attaining F_1-macro scores comparable to those from previous settings. However, neither of these experiments outperformed the previously achieved scores, indicating possible areas for future work.

The most effective model overall was microsoft/deberta-base-v2-xlarge[33] 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 4, with the small gap between validation and test results indicating the model's ability to generalise effectively. Table 5 displays the confusion matrix, depicting actual labels on the y-axis and predicted labels on the x-axis. We observed greater F_1 score performance for fake experts, anecdote, conspiracy theory and ad hominem. In contrast, false equivalence and slothful induction exhibited the lowest F_1 scores.

Comparing FLICC model to Alhindi et al.[26] and Jin et al.[25]

Although the comparison is not straightforward, both[25] and[26] developed climate change fallacy datasets, training machine learning models with similar numbers of fallacies (13 and 9 respectively). They reported overall F_1 scores of 0.21 and 0.29 for their climate datasets in their best round of experiments, whereas we achieved an F_1 score 0.73, indicating a performance improvement by a factor of 2.5 to 3.5. However, a direct comparison between these studies and our results are difficult as we do not share the same set of fallacies. But, Table 6

	Gemini			GPT-4		
	P	R	F_1	P	R	F_1
Ad hominem	0.00	0.00	0.00	0.86	0.32	0.47
Anecdote	0.00	0.00	0.00	0.46	0.50	0.48
Cherry picking	0.45	0.29	0.35	0.20	0.10	0.13
Conspiracy theory	0.42	0.86	0.57	0.53	0.91	0.67
Fake experts	0.00	0.00	0.00	0.75	0.86	0.80
False choice	0.50	0.14	0.22	1.00	0.14	0.25
False equivalence	0.00	0.00	0.00	0.20	0.12	0.15
Impossible expectations	0.00	0.00	0.00	0.17	0.05	0.07
Misrepresentation	0.14	0.09	0.11	0.31	0.23	0.26
Oversimplification	0.13	1.00	0.22	0.14	0.60	0.23
Single cause	0.00	0.00	0.00	0.36	0.25	0.30
Slothful induction	0.00	0.00	0.00	0.12	0.08	0.10
Accuracy			0.20			0.32
Macro avg	0.13	0.18	0.11	0.39	0.32	0.30
Weighted avg	0.13	0.20	0.12	0.40	0.32	0.31

Table 2. Fallacy classification results for Google's Gemini and OpenAi's GPT-4 models. For each class, we report precision (P), recall (R), and F_1 score.

	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 3. F_1 macro scores, highlighted cells indicate the best model parameter combination for each model. Best model overall was microsoft/deberta-base-v2-xlarge, learning rate 1.0e-5, gamma 4, weight decay 0.01 fine-tuned over 15 epochs.

provides a summary of the results for the shared fallacies between the scores obtained by[25] and[26] using their respective models on their datasets, and our model's performance on our dataset.

Discussion

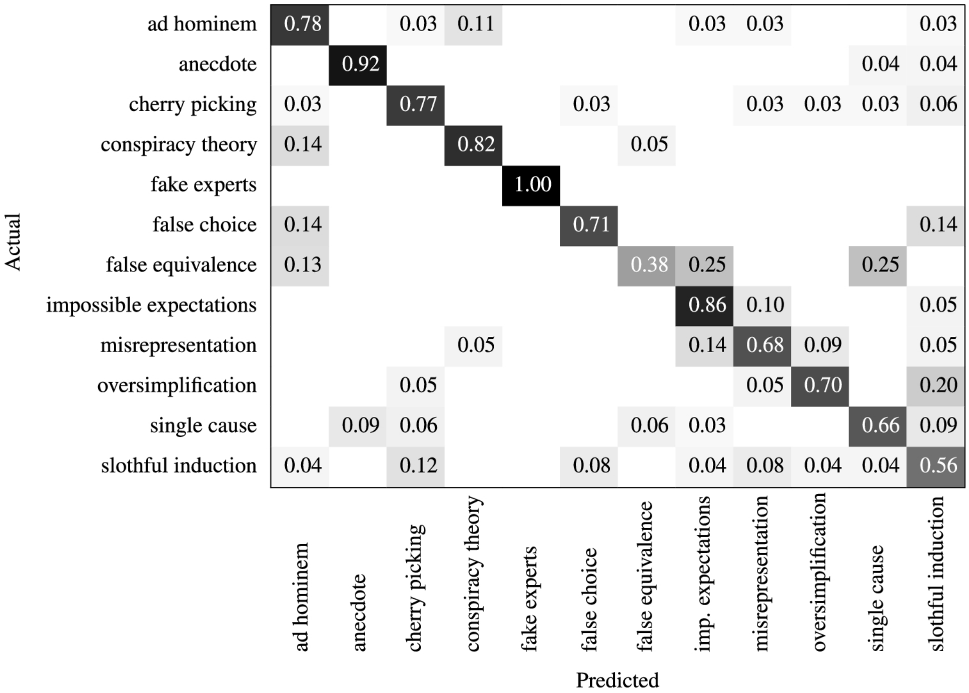
In this study, we developed a model for classifying logical fallacies in climate misinformation. Our model performed well in classifying a dozen fallacies, showing significant improvement on previous efforts. The Deberta model also showed better results than those obtained from Gemini-pro and GPT-4 models. 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 exhibited lower performance with certain fallacies compared to others, with the false equivalence fallacy displaying the lowest performance, likely due to 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 potential contributor to the difficulty in detecting slothful induction was the conceptual overlap between slothful induction and cherry picking. Both fallacies involve 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 a conclusion without citing evidence[15]. Another factor to consider in analysing the poor performance of slothful induction, as illustrated in Fig. 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 is more evenly distributed across all claim topics.

Another source of difficulty are texts that contain multiple fallacies. It is common that climate misinformation incorporates several elements in a single item. An example is making a content claim such as "a cooling sun

	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 4. FLICC model fallacy classification report. For each class, we report precision (P), recall (R), F_1 score for validation and test partitions.



	ad hominem	anecdote	cherry picking	conspiracy theory	fake experts	false choice	false equivalence	imp. expectations	misrepresentation	oversimplification	single cause	slothful induction
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 5. Normalised confusion matrix, actual labels on y-axis, predicted labels on x-axis.

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[23]. Further, some texts may include a single claim that nevertheless contains multiple fallacies. For example, the claim that "there's no evidence that CO_2 drove temperature over the last 400,000 years" commits slothful induction by ignoring all the evidence for CO_2 warming as well as false choice by demanding that either CO_2 drives temperature or temperature drives CO_2[15].

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

Alhindi et al.[26]	max. F_1	F_1	FLICC
Causal oversimplification	0.53	**0.72**	Aingle cause
Cherry picking	0.43	**0.77**	Cherry picking
Irrelevant authority	0.30	**1.00**	Fake experts
Jin et al.[25]	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 6. Summary of F_1 scores for comparable labels (fallacies). On the left side we have labels from Alhindi et al.[26] and Jin et al.[25]. On the right side, the FLICC model labels.

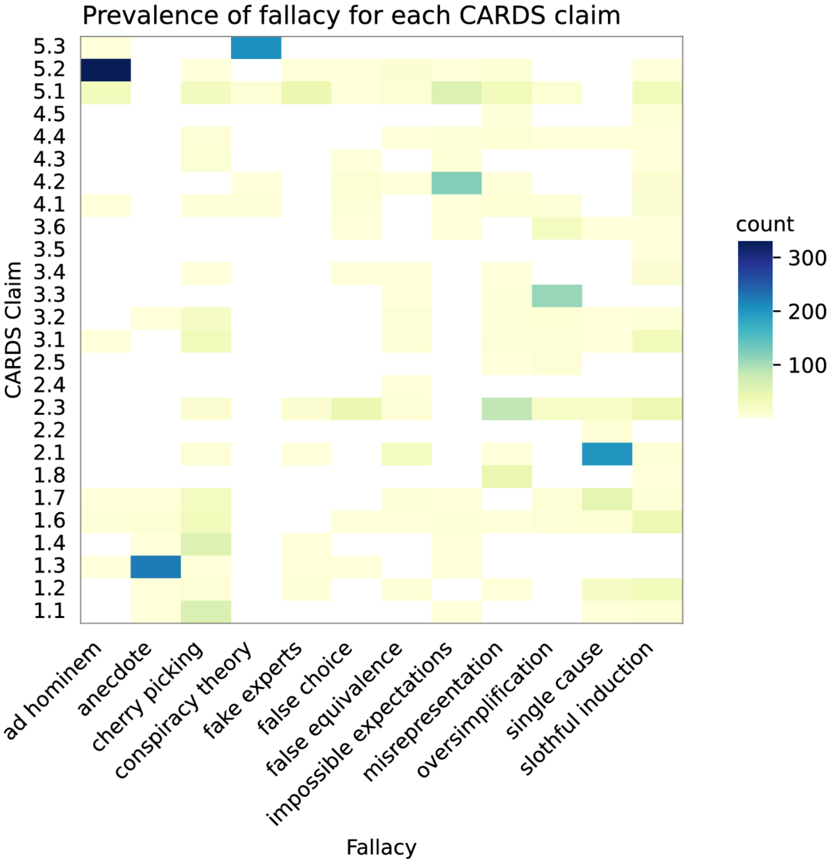


Fig. 3. Map of fallacies across different CARDS claims.

accessible avenue involves creating an automated ReAct agent[34] that we can further optimise using evolutionary computation techniques[35]. A more sustainable, long-term approach might involve fine-tuning a LLM[36,37].

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[29]. 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[23,24]. Potentially, a model that can detect both content claims in climate misinformation and fallacies could generate corrections that adhere to the fact-myth-fallacy structure recommended by psychological research[11].

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[14,15]. There is a dearth of 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 optimal response strategies.

Our research also demonstrates the contribution that critical thinking can offer to computer science research. Our work is based on manual deconstruction of contrarian climate claims, a necessary step as misleading claims can be based on unstated assumptions or hidden premises[14]. Indeed an analysis of contrarian claims about climate change found that the majority of claims contained hidden premises which committed reasoning fallacies[15].

Another important consideration when assessing potential misinformation is the use of factual statements to paint a misleading impression by withholding relevant information, a technique known as paltering or cherry picking[13,38]. We leveraged advancements in critical thinking research, using manually deconstructed misinformation claims, to develop a curated training dataset of fallacy examples. This is not to say that all statements about climate change can be unambiguously classified as true and false, and measures for determining which statements are fact-checkable and which are not are required. Nevertheless, there exist many incontrovertible facts and conversely, misleading statements that contain clearly misleading fallacies, that are rightfully subject to flagging as misleading content[39].

The development of interventions that detect and counter misinformation also raises ethical questions, as such efforts can potentially be exploited by bad faith actors such as repressive governments seeking to suppress free speech[40,41]. Because of these concerns, transparency and clarity of purpose are essential when developing misinformation interventions. In the case of our fallacy detection model, its purpose is not intended to facilitate censorship but to facilitate explanations of reasoning fallacies used in misinformation, thus building the public's critical thinking skills. For example, one application that is currently under development is a tool using a large language model to generate automated responses to misinformation that incorporate explanations of misleading fallacies[42]

Another ethical consideration is the impact that misinformation has to undermine democracy and impinge on the public's right to be accurately informed[39,43]. Because of these and other harmful impacts, misinformation should not remain unchallenged[44]. Interventions that strengthen the public's capacity to discern factual information from misinformation upholds democracy and bolsters people's freedom from being misinformed. In particular, technique-based interventions which our fallacy-detection model is designed to support increase the public's ability to spot manipulation techniques. Past work on boosting people's metacognition, defined as insight into the accuracy of knowledge and beliefs[45], by warning them about the misleading threat of specific logical fallacies, has been shown to be effective in neutralizing climate misinformation across the political spectrum[2].

The interaction between psychological and computer science research illustrates the value of the technocognitive approach to misinformation research. Inevitably, technological solutions will 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 from psychological and critical thinking work[2,8,14,23], and any output from such models should be informed by psychological research.

Methods

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[23] training set, the climate datasets from[26] and[25], and the test set of climate tweets from[24]. In order to more reliably identify dominant fallacies in text, we employed the critical thinking methodology from[14] to deconstruct difficult examples. Table 7 shows a selection of sample deconstructions of the most common combinations of CARDS claims and FLICC fallacies.

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[46] to construct a similarity matrix, utilising cosine similarity (Eq. 3) 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; in such cases, we retained the most representative of these samples.

$$\cos\varphi = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{3}$$

$$d\left(p,q\right) = \sqrt{p\cdot p - 2(p\cdot q) + q\cdot q} \tag{4}$$

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 (Eq. 4) 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[47], 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 did not 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 Fig. 3. It showed a much broader distribution of fallacies within each CARDS claim than found in[15]. This indicated that contrarian arguments could take various forms featuring different fallacies, and that merely

Misinformation example	Claim	Deconstruction	Fallacy explanation
"I'll believe in climate change when elitists stop building mansions on the coast"	5.2	P1: Climate advocates argue for climate action P2: Climate advocates' actions are inconsistent with their arguments HP: If climate advocates are inconsistent, their arguments must be invalid C: Arguments for climate action can be disregarded	HP commits ad hominem, criticising climate advocates rather than their arguments
"Global Warming? Tell that to the southern districts that woke up to negative 10 degrees this morning"	1.3	P1: Cold weather events are occuring. HP: If global warming was happening, we wouldn't experience cold events C: Global warming is not happening	P1 commits anecdote, using isolated incidents limited in time and place to make conclusions about global warming
"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 hasn't changed much in the short term, then it's 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
"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
"More than 31,000 American scientists signed a statement saying they disagree with alarmist predictions"	5.1	P1: A large number of scientists disagree with human-caused global warming HP: Scientists are experts on climate change regardless of their field of expertise C: There's no scientific consensus on human-caused global warming	HP commits fake experts. While the signers of the global warming petition project are scientists, almost all of them don't possess climate expertise
"Who denies that CO2 lags temperature in the ice core data by as much as 800 years and hence is a product of climate change not a cause?"	2.3	P1: CO2 lagged temperature in the past. HP: If temperature affects CO2, then CO2 cannot affect temperature C: CO2 does not drive temperature	HP presents a false choice between CO2 causing warming or warming causing CO2, while both are true
"Tuesday is Earth Day, the calendar's High Holy Day of Green theology. With each passing year, environmentalism more clearly assumes the trappings of a secular religion"	5.2	P1: The climate change movement have some trait in common with religion HP: A movement that has any traits in common with a religion is a religion C: The climate change movement is a religion	HP commits false equivalence, making superficial comparisons between the climate movement and religion, when climate science is based on empirical evidence, not faith
"A 40% reduction in US emissions would have no measurable impact on atmospheric CO2 increase"	4.2	P1: A single policy would have a negligible impact HP: If a single policy doesn't solve global warming, then it is not worth implementing C: We should not have the policy	HP commits impossible expectations. A single policy cannot solve climate change by itself. We need global cooperation to solve climate change
"CO2 is incapable of causing climatic warming. 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 isn't 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)
"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. The full picture shows that negative impacts outweigh benefits
"At the current sea-level-equivalent ice-loss rate of 0.05 millimeters per year, it would take a full millennium to raise global sea level by just 5 cm, and it would take fully 20,000 years to raise it a single meter"	1.6	P1: Sea level is rising at a modest rate HP: The rate of sea level rise won't increase in the future C: Future sea level rise will not be large	HP commits slothful induction, ignoring that sea level rise is accelerating and predicted to increase in the future
"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

Table 7. Deconstructions examples representing 12 fallacies.

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.

Training a model to detect fallacies

Model selection

Classifying fallacies, especially when they revolve around a singular subject such as climate change, poses a significant challenge. Ref.[25] 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 Fig. 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[32] might offer a promising initial solution to our problem. LoRa brings several advantages in terms of storage and

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	1796	457	256	2509

Table 8. Fallacy types and their number of samples on each partition in the FLICC dataset.

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 classification problem.

In order to test our hypothesis, we evaluated all accessible models within HuggingFace's Parameter-Efficient Fine-Tuning (PEFT) library (https://github.com/huggingface/peft) 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*.

Experimental setup

We employed the PyTorch (https://pytorch.org) framework and HuggingFace (https://huggingface.co) libraries for our experiments, conducting an iterative analysis to optimise the configuration at each experimental stage. Our dataset was partitioned into train, validation, and test sets as illustrated in Table 8. 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 F_1-macro score, considering the imbalanced nature of our dataset.

We examined the best learning rates within 1.0e−5, 5.0e−5 and 1.0e−4. We set the batch size to 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[48] 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.

Data availibility

The dataset and the codes to train our model are available in the GitHub repository https://www.github.com/fzanart/FLICC. The data and code are licensed under the MIT License, allowing for reuse and adaptation with proper attribution. For any questions or issues, please email francisco.zanartu@unimelb.edu.au.

Received: 24 June 2024; Accepted: 10 October 2024
Published online: 12 November 2024

References

1. Ranney, M. A. & Clark, D. Climate change conceptual change: Scientific information can transform attitudes. *Top. Cogn. Sci.* **8**, 49–75 (2016).
2. Cook, J., Lewandowsky, S. & Ecker, U. K. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One* **12**, e0175799 (2017).
3. Van der Linden, S., Leiserowitz, A., Rosenthal, S. & Maibach, E. Inoculating the public against misinformation about climate change. *Global Chall.* **1**, 1600008 (2017).
4. Geiger, N. & Swim, J. K. Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *J. Environ. Psychol.* **47**, 79–90 (2016).
5. Kozyreva, A. *et al.* Toolbox of interventions against online misinformation and manipulation. (2022).
6. Schmid, P. & Betsch, C. Effective strategies for rebutting science denialism in public discussions. *Nat. Hum. Behav.* **3**, 931–939 (2019).
7. Banas, J. A. & Miller, G. Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Hum. Commun. Res.* **39**, 184–207 (2013).

8. Vraga, E. K., Kim, S. C., Cook, J. & Bode, L. Testing the effectiveness of correction placement and type on instagram. *Int. J. Press/Polit.* **25**, 632–652 (2020).
9. McCright, A. M., Charters, M., Dentzman, K. & Dietz, T. Examining the effectiveness of climate change frames in the face of a climate change denial counter-frame. *Top. Cogn. Sci.* **8**, 76–97 (2016).
10. Lewandowsky, S., Cook, J. & Ecker, U. K. Letting the gorilla emerge from the mist: Getting past post-truth. *J. Appl. Res. Mem. Cogn.* **6**, 418–424 (2017).
11. Lewandowsky, S. *et al.* The debunking handbook 2020 (2020).
12. Diethelm, P. & McKee, M. Denialism: what is it and how should scientists respond?. *Eur. J. Public Health* **19**, 2–4 (2009).
13. Cook, J. Deconstructing climate science denial. In *Research Handbook on Communicating Climate Change* 62–78 (2020).
14. Cook, J., Ellerton, P. & Kinkead, D. Deconstructing climate misinformation to identify reasoning errors. *Environ. Res. Lett.* **13**, 024018 (2018).
15. Flack, R. *et al.* Identifying reasoning fallacies in a comprehensive taxonomy of contrarian claims about climate change. *Environ. Commun.* (2024).
16. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science.* **359**, 1146–1151 (2018).
17. Ecker, U. K., Lewandowsky, S. & Tang, D. T. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Mem. Cogn.* **38**, 1087–1100 (2010).
18. Cook, J. Understanding and countering climate science denial. *J. Proc. R. Soc. N.S.W.* **150**, 207–219 (2017).
19. Hassan, N. *et al.* The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium* (Citeseer, 2015).
20. Boussalis, C. & Coan, T. G. Text-mining the signals of climate change doubt. *Glob. Environ. Change* **36**, 89–100 (2016).
21. Farrell, J. Corporate funding and ideological polarization about climate change. *Proc. Natl. Acad. Sci.* **113**, 92–97 (2016).
22. Stecula, D. A. & Merkley, E. Framing climate change: Economics, ideology, and uncertainty in American news media content from 1988 to 2014. *Front. Commun.* **4**, 6 (2019).
23. Coan, T. G., Boussalis, C., Cook, J. & Nanko, M. O. Computer-assisted classification of contrarian claims about climate change. *Sci. Rep.* **11**, 22320 (2021).
24. Rojas, C. *et al.* Augmented cards: A machine learning approach to identifying triggers of climate change misinformation on twitter (2024).
25. Jin, Z. *et al.* Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022* 7180–7198 (eds. Goldberg, Y., Kozareva, Z. & Zhang, Y.). https://doi.org/10.18653/v1/2022.findings-emnlp.532 (Association for Computational Linguistics, 2022).
26. Alhindi, T., Chakrabarty, T., Musi, E. & Muresan, S. Multitask instruction-based prompting for fallacy recognition (2023). arXiv:2301.09992.
27. Lewandowsky, S., Ecker, U. K. & Cook, J. Beyond misinformation: Understanding and coping with the "post-truth" era. *J. Appl. Res. Mem. Cogn.* **6**, 353–369 (2017).
28. Roozenbeek, J. & Van der Linden, S. Fake news game confers psychological resistance against online misinformation. *Palgrave Commun.* **5**, 1–10 (2019).
29. Hopkins, K. L. *et al.* Co-designing a mobile-based game to improve misinformation resistance and vaccine knowledge in uganda, kenya, and rwanda. *J. Health Commun.* (2023).
30. Team, G. *et al.* Gemini: A family of highly capable multimodal models (2024). arXiv:2312.11805.
31. OpenAI *et al.* Gpt-4 technical report (2024). arXiv:2303.08774.
32. Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models (2021). arXiv:2106.09685.
33. He, P., Liu, X., Gao, J. & Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations* (2021).
34. Yao, S. *et al.* React: Synergizing reasoning and acting in language models (2023). arXiv:2210.03629.
35. Fernando, C., Banarse, D., Michalewski, H., Osindero, S. & Rocktäschel, T. Promptbreeder: Self-referential self-improvement via prompt evolution (2023). arXiv:2309.16797.
36. An, S. *et al.* Learning from mistakes makes llm better reasoner (2023). arXiv:2310.20689.
37. Huang, J. *et al.* Large language models cannot self-correct reasoning yet (2023). arXiv:2310.01798.
38. Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I. & Schweitzer, M. E. Artful paltering: The risks and rewards of using truthful statements to mislead others. *J. Pers. Soc. Psychol.* **112**, 456 (2017).
39. Ecker, U. et al. Misinformation poses a bigger threat to democracy than you might think. *Nature* **630**, 29–32 (2024).
40. Riemer, K. & Peter, S. Algorithmic audiencing: Why we need to rethink free speech on social media. *J. Inf. Technol.* **36**, 409–426 (2021).
41. Warf, B. Geographies of global internet censorship. *GeoJournal* **76**, 1–23 (2011).
42. Zanartu, F., Otmakhova, Y., Cook, J. & Fermann, L. Generative debunking of climate misinformation. arXiv preprint arXiv:2407.05599 (2024).
43. Lewandowsky, S. et al. Liars know they are lying: Differentiating disinformation from disagreement. *Humanit. Soc. Sci. Commun.* **11**, 1–14 (2024).
44. Ecker, U. K. *et al.* Why misinformation must not be ignored. *Am. Psychol.* (2024).
45. Fischer, H. & Fleming, S. Why metacognition matters in politically contested domains. *Trends Cogn. Sci.* (2024).
46. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding (2019). arXiv:1810.04805.
47. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17 (2008).
48. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection (2018). arXiv:1708.02002.

Author contributions

F.Z., J.C., M.W. and J.G. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.