Evaluating LLMs for Detecting Climate Misinformation: How Aligned are LLMs with Expert Classification of False or Misleading Claims about Climate Change on Social Media?

Anonymous ACL submission

Abstract

Online dis/misinformation is a social problem that has the potential to be substantially aggravated by the development of Large Language Models (LLMs). In this study we evaluate the potential for LLMs to be part of the solution for mitigating online dis/misinformation rather than the problem. Employing a public expert annotated dataset and a curated sample of social media content we evaluate the performance of GPT-4 and fine-tuned GPT-3.5-turbo on climate misinformation classification task, comparing them to existing climate-focused computer-assisted tools and expert assessments. Results show that fine-tuned GPT-3.5-turbo outperforms GPT-4 and a BERT-based model for classifying climate misinformation and functions at the equivalency of climate change experts with over 20 years of experience in climate communication. These findings highlight the potential of fine-tuned LLMs in 1) facilitating civil society organizations with limited technical expertise to engage in a range of governance tasks with respect to climate misinformation, and 2) encourage further exploration of AI-driven solutions for detecting climate dis-/misinformation.

1 Introduction

When considering public understanding of, mobilization about, and support for climate change science and policy, people rely on information from mediated sources such as online or offline media rather than directly from scientists (Scheufele, 2014). The reliance on such sources provides an opportunity for false or misleading claims about climate change to compete with accurate ones and influence public opinion and policy discourse which seriously hampers climate mitigation efforts (Allgaier, 2019; Gounaridis and Newell, 2024; IPCC, 2022; Lewandowsky, 2021; Treen et al., 2020).

Advances in Large Language Models (LLMs) could also contribute to the exacerbation and diffusion of climate misinformation if exploited by malicious actors such as authoritative regimes. LLMs are capable of generating high volumes of persuasive, deceptive, and human-like content full of misinformation that promotes climate change denial and skepticism (Fore et al., 2024; Zhang et al., 2024; CAAD, 2024; Ellison and Hugh, 2024) making it more difficult for humans to detect LLM-generated dis/misinformation in news context especially when circulating on social media (Kreps et al., 2022; Marlow et al., 2020).

The major knowledge gaps about the nature of climate change information on digital platforms, combined with inconsistent and ineffective content moderation policies across platforms (CAAD, 2024; CCDH, 2021; Romero-Vicente, 2023), continues to seriously hamper the effective mitigation of climate change misinformation social media. A key element of this problem is inadequate identification and classification tools that have the necessary scale, scope, and (especially) technical expertise required to evaluate the veracity of claims about climate change circulating online (Coan et al., 2021; Vu et al., 2023).

In response, researchers have attempted to develop a variety of computer-assisted classification tools to identify and respond to false or misleading claims about climate change (Coan et al., 2021; Leippold et al., 2024). However, the generalizability of these detectors is less robust on types and sources of text beyond what they are trained on, demonstrating how easy it is for malicious actors to overcome the detection by these algorithms (Stiff and Johansson, 2022).

To address these issues, we take a human-centered approach into evaluating the generalizability of LLMs in detecting false or misleading claims about climate change present in articles sourced from low credible news sources that are circulating on social media. We incorporate the assessment of two experts with over 20 years of experience

in climate communication into the evaluation of the capability of these LLMs in classifying climate misinformation. By pairing a quantitative evaluation with expert assessment of LLMs, we overcome some of the limitations related to the validity, quality, and diversity of automated benchmarks that are commonly used by the AI and NLP communities when performing automated evaluation of LLMs(Gehrmann et al., 2023; Xiao et al., 2024).

Within this context, we explore in this work four research questions: (1) how well does GPT-4 classify false or misleading claims about climate change in zero-shot classification? (2) does fine-tuning a smaller LLM (GPT-3.5-turbo) yield better performance in classifying climate misinformation as compared to GPT-4, (3) how do these models compare to existing computer-assisted approaches, and (4) and how aligned are they with domain expert classification of climate change claims circulating on social media? The results of our study provide the following contributions:

- We identify GPT-4's inferior performance as compared to a BERT-based computer assisted tool (CARDS) (Coan et al., 2021) for classifying false or misleading claims about climate change on an expert annotated dataset.
- We demonstrate the superior performance of fine-tuning GPT-3.5-turbo, as compared to GPT-4 and CARDS (Coan et al., 2021), for classifying false or misleading claims commonly found in climate skeptic and contrarian blogs as annotated by climate change experts.
- We illustrate GPT-3.5-turbo's strong and extensible capability for classifying false or misleading claims about climate change commonly found in social media with the approximate reliability as two senior climate change communication experts with over 20 years of experience.

Through our findings, we hope to contribute to the collective efforts aimed at assessing the efficacy of LLMs in detecting and guardrailing against climate dis/misinformation. In addition, we encourage AI researchers, designers, and developers to reflect on our approach and findings as they consider deploying LLMs for content moderation or as part of automated evaluation workflows for identifying false or misleading claims about climate change in either human or AI generated content.

2 Related Work

The potential abuse of LLMs by malign actors to generate misinformation that shapes the information environment and public opinion has become evident across a range of domains, from politics to healthcare, and including climate change (Yang and Menczer, 2024; Marlow et al., 2020; Ferrara et al., 2020; Akhtar et al., 2023; De Angelis et al., 2023). However, detecting climate dis/misinformation generated by humans is difficult, let alone content generated by LLMs (Chen and Shu, 2023). In an effort to combat climate dis/misinformation, researchers have introduced tools that leverage LLMs to identify false claims about climate change from skeptic and contrarian sources, fact-check these claims, and even generate factual and in-depth answers to climate related questions (Coan et al., 2021; Leippold et al., 2024; Thulke et al., 2024; Mullappilly et al., 2023). Despite these efforts, researchers have questioned the generalizability of dis/misinformation detection tools and have identified vulnerabilities that enable malicious content to bypass them (Stiff and Johansson, 2022). One proposed solution to this issue is to establish benchmarks to evaluate the capability of LLMs in classifying and detecting climate related content.

Researchers have constructed several datasets relevant to climate change research, but not necessarily focused on benchmarking for climate dis-/misinformation detection, to evaluate LLMs for climate-related classification tasks such as: predicting sentence relevance to climate change (Leippold et al., 2024), stance detection in support or opposition toward climate change prevention (Vaid et al., 2022), and fact-checking scientific information related to climate science (Laud et al., 2023; Pirozelli et al., 2023).

Still, such benchmarks seem to be suffering from limitations in terms of the diversity and complexity of examples used to rigorously evaluate the input and generated text of the models for different downstream applications (Liang et al., 2022). It is essential, therefore, to involve stakeholders such as climate change experts in the design, development, and validation for many of these tools and datasets, as their input and feedback not only refines the quality of the benchmark datasets, but also contributes to model enhancements in terms of handling climate misinformation (Stiennon et al., 2020; Zhou and Xu, 2020; Ouyang et al., 2022; Christiano et al., 2017). Through such human-centered approach,

stakeholders can identify in-depth criteria for what these models are being evaluated on and propose edge cases that impose constraints on the model performance before it gets deployed(Xiao et al., 2024).

3 Data

Our study employs two datasets: 1) a public dataset of false or misleading claims about climate change sourced (see section 3.1) used to train Climate Change Denial and Skepticism (CARDS) model (Coan et al., 2021) and 2) a curated sample of the most engaging articles and blog posts (i.e., based on the number of likes, comments, shares) about climate change on Facebook and X (i.e., Twitter) from right-biased, questionable, and low credible sources.

3.1 CARDS (benchmarking) dataset

The curated dataset to train the CARDS model contains paragraphs from 53 contrarian and skeptical domains about climate change spanning the years 1998 to 2020 and their corresponding claims based on annotations from the authors who are experts in climate research (Coan et al., 2021). The dataset is organized by the authors of CARDS into a taxonomy of super claims and sub-claims that represent the primary arguments employed by climate denialists and skeptics (see Appendix A.2). False and misleading claims in this taxonomy are grouped into five main categories: (1) global warming is not happening, (2) humans greenhouse gases are not causing global warming, (3) climate impacts are not bad, (4) climate solutions won't work, and (5) the climate movement and/or science are unreliable.

The dataset is split into training set (N=23,436), validation set (N=2,605), and test set (N=2,904) each containing paragraphs from sources publishing climate misinformation and their corresponding annotated claims by domain experts. Our decision to use this dataset is based on the (1) breadth of this dataset on climate misinformation and (2) the taxonomy of claims developed and validated by experts in climate-research sourced from two decades worth of skeptic and denialist content (Coan et al., 2021). We leveraged this dataset because of the granularity it offers at the paragraph level in the training, validation, and test datasets which which allows benchmarking at different levels of analysis (e.g. paragraph level to simulate social media posts or aggregated at to the article level). The availability of the coding manuals used to annotate the claims also enables us to craft prompts using the instructions in these manuals for GPT-4 to classify false and misleading claims about climate change (as later described in section 4.1).

Articles in the CARDS dataset were sourced from 20 prominent conservative think tanks (CTTs) and 33 blogs. Every article was split into multiple paragraphs each having a sequence length of 256 characters. All paragraphs in the dataset are in English language and were cleaned from URLs, scholarly citations, and non-English content. In addition, claims were annotated by the authors of CARDS based on the taxonomy of claims they developed (Coan et al., 2021) and included in the Appendix (see A.2). Claims in the CARDS data are formatted as strings that combine the super-claim and sub-claim into a single label separated by an underscore (e.g., "5_1") referring to the super-claim and sub-claim, respectively.

3.2 Social Media dataset

Scraping data

Using an API from NewsWhip[1], a social media analytics platform, we retrieved the URLs for the daily 5000 most engaging (e.g. likes, shares, comments) English-language articles discussing climate change that were published from American domains on Facebook and X between January and December 2022, inclusive. A list of relevant keywords compiled by a climate change communication expert was used to query the NewsWhip API. A full list of the keywords can be found in the Appendix (see Appendix A.1). Next, we scraped the text and metadata (e.g., publication date) from all the URLs retrieved from NewsWhip using a custom web scraper in Python that leverages Newspaper3K[2] and BeautifulSoup[3] libraries.

We ensured the scraped articles were centered on climate change, and not merely mentioning the topic in passing, by filtering the corpus using the same list of keywords used for the API search (see Appendix A.1). The keyword filter was applied to the headline and first 250 words of each scraped article, resulting in a corpus of 829,827 articles with climate change and published on Facebook or X in 2022.

[1] https://www.newswhip.com/

[2] https://newspaper.readthedocs.io/en/latest/

[3] https://pypi.org/project/beautifulsoup4/

Domain credibility

We filtered our corpus by the domain credibility of the publisher as means to curate a test dataset that contained a sufficient number of false or misleading claims about climate change circulating on social media,.To determine the domain level credibility of each scraped article, we relied on a combination of the Media Bias Fact Check (MBFC) categories[4] and NewsGuard Trust score[5]. MBFC categorizes news sources in one of nine bias categories: least biased, left bias, left-center bias, right-center bias, right bias, conspiracy-pseudoscience, questionable sources, pro-science, and satire. Similarly, NewsGuard Trust score is a reliability rating between 0 and 100 that is assigned by journalists and editors to news websites based on journalistic and apolitical criteria such as credibility and transparency. A NewsGuard score of 60 or below indicate an untrustworthy news source.

Employing these two resources, we appended the MBFC category and the NewsGuard score to all articles in our corpus with domains matching those in the two datasets. We then selected all articles from MBFC right-bias, conspiracy-pseudoscience, and questionable categories and/or had a Newsguard score of 60 or below as articles that are most likely to contain false or misleading information. Out of the 829,927 articles published about climate change in 2022 on Facebook and X, 71,175 (8.6%) were classified as originating from right-bias, conspiracy-pseudoscience, and questionable domains with low credibility[6].

4 Methodology

This section illustrates how the CARDS dataset described in section 3.1 is leveraged to (1) prompt and evaluate GPT-4 (section 4.1), and (2) fine-tune GPT-3.5-turbo to classify false or misleading claims about climate change (section 4.2). In addition, we describe how we applied the fine-tuned and CARDS models on a sample of articles from social media to further test and validate the models' ability to classify climate misinformation from other sources beyond contrarian content from conservative think tanks and contrarian blogs (section 4.3). Finally, in section 4.4, we outline how two climate communication experts annotated the claims from a sample of paragraphs sourced from social media and how these annotations were used to measure the alignment between the LLMs and expert classification of false or misleading claims about climate change.

4.1 Evaluating GPT-4 on CARDS

To assess how well GPT-4 performs in classifying false or misleading claims about climate change, we apply zero-shot classification technique to classify paragraphs about climate change, sourced from articles published by conservative think tanks and blog posts, that are part of the CARDS test dataset. This test dataset serves as a baseline for comparing the generated claims by GPT-4 and those already annotated by climate-research experts in the CARDS test dataset for each paragraph.

To prompt GPT-4 to classify false or misleading claims about climate change, we crafted a system and user prompts based on the instructions derived from the coding manual used to train annotators for labeling the training and testing datasets that were subsequently used to train the RoBERTa$_{large}$ CARDS model (Coan et al., 2021). The system and user prompt structure are outlined in Appendix A.4 and A.5.

The system prompt includes an overview of the task and a coding rubric structured in JSON format containing an identified, code, and claim description keys for each claim and sub-claim as described in the CARDS supplimentary information(Coan et al., 2021). The model is prompted to output either (1) "no claim" when no false or mis-leading claim is present in a paragraph or (2) one of 16 labels[7] corresponding to false or misleading claims outlined in the taxonomy of claims for the CARDS (see Appendix A.2)

The user prompt was structured in a question and answer format asking the model about the claim to which an excerpt of text belongs to (as shown in Appendix A.5). The phrasing of the question was also derived from the CARDS coding manual. We crafted the prompts to be consistent with the CARDS coding manual so we have a one-to-one comparison of the performance between the generated claim labels by GPT-4 and the annotated claims by expert coders. Next, we validated the prompts on the CARDS validation dataset to confirm that the model is compliant with the system

[4] https://www.mediabiasfactcheck.com

[5] https://www.newsguardtech.com/solutions/newsguard/

[6] Check Appendix A.3 for details about the most prevalent domains in our dataset

[7] As described in section 3.1 each label combines a super-claim and its corresponding sub-claim delimited by "_".

and user prompts and that the generated claims are in the same format as in the CARDs datasets.

To evaluate GPT-4's performance in classifying claims about climate change, we leveraged the test set (N=2,904) described in section 3.1. Using OpenAI library, we sent individual requests for each paragraph in the test dataset to GPT-4-0125 via the chat completion endpoint. Each request includes the system and user prompts outlined in Appendix A.4 and A.5. We requested the response format of the model to be in "json/object" which guarantees the returned message that the model generates is in valid JSON format. This enables easier handling and parsing of the returned response from the API endpoint. Also, to avoid GPT-4 from generating preamble messages as part of the response, we configured the temperature to 0. This ensures a more deterministic and concise responses that include only the label of the claim to which a paragraph belongs to.

To compare the generated claims using GPT-4 to those present in the test dataset we focused only at the super-claim level as a proxy to assess the model's performance on the CARDS dataset. Accordingly, we split the claim labels in the generated claims and those in the test dataset based on the under score delimiter. Only the number to the left of the underscore ("_") delimiter was extracted for model evaluation and comparison as it represents the super-claim. We elaborate on our findings from benchmarking the performance of GPT-4 with respect to the test dataset of the CARDS model in section 5.1 of the results section.

Next, we describe our attempt to fine-tune GPT-3.5-turbo on CARDS datasets.

4.2 Fine-tuning GPT-3.5-turbo on CARDS

Although identifying false or misleading claims about climate change using zero-shot classification is feasible, it is not as scalable and extensible approach to detecting misinformation as fine-tuning. For instance, the length and amount of contextual details included in the prompts, along with the need to continuously update, test, and validate these prompts (especially when requesting the model to classify new types of emerging claims about climate change) makes zero-shot prompting a less scalable approach when compared to fine-tuning. In contrast, fine-tuning enhances the ability of general purpose LLMs to be more optimal in their adaptability to domain-specific tasks such as identifying emerging types of claims about climate change by including a few examples of these claims in the training data. In addition, fine-tuned LLMs that accurately classify false or misleading information about climate change have the potential to be integrated into a variety of governance tools such as those used for content moderation (Leippold et al., 2024).

For these reasons, we chose to fine-tune GPT-3.5-turbo to classify false or misleading claims and evaluate the alignment of the classified claims with respect to the claims classified by GPT-4 and those originally annotated by CARDS climate experts in the CARDS test set. Similar to section 4.1, we rely on the training (N=23,436) and validation (N=2,605) sets from CARDS to fine-tune GPT-3.5-turbo and evaluate its performance on classifying false or misleading claims about climate change in the test set. Employing OpenAI's Python library, we fine-tune GPT-3.5-turbo for 3 epochs with a batch size of 8. The number of epochs and batch size were selected to closely reflect the training hyperparamters for RoBERTa$_{large}$ CARDS model[8].

Because GPT-3.5-turbo is an instruction-based model, we had to structure the CARDS training and validation datasets in a list of chat message dictionaries as recommended by OpenAI[9]. Each dictionary includes a system and user messages as shown in Appendix A.6. The system message includes the system prompt explaining to the model details about classifying false or misleading claims about climate change. The content of the user message includes the paragraph text from the training or validation sets. Finally, the content of the assistant message is left empty to prompt the model to generate a claim label that corresponds to the paragraph in the user message.

All compiled dictionaries of system and user messages were grouped in a JSONL file format to comply with OpenAI fine-tuning requirements. Accordingly, GPT-3.5-turbo was fine-tuned on 13,330,863 tokens for 8,789 steps and costed $106.65 USD.

To assess the performance of the fine-tuned model with respect to the RoBERTa$_{large}$ CARDS model, we used the fine-tuned model to generate the corresponding claim label for each paragraph in the CARDS test set. To send these requests to the chat completion endpoint, we also used OpenAI

[8] RoBERTa$_{large}$ CARDS model was trained on 3 epochs and batch size of 6

[9] https://platform.openai.com/docs/guides/fine-tuning/example-format

Python library and formatted the requests as chat messages similar to what we have already done for zero-shot classification in section 4.1. We elaborate on the findings from our assessment in section 5.1 of the results.

Next, we describe how the fine-tuned GPT-3.5-turbo and the RoBERTa$_{large}$ CARDS models were applied to classify false or misleading claims about climate change in paragraphs from most engaging and low-credibility articles on social media about climate change.

4.3 Classifying claims in social media data

Employing the social media dataset described in section 3.2, we utilize the fine-tuned GPT-3.5-turbo to classify claims underpinning misinformation about climate change at the paragraph level similar to CARDS. The fine-tuned model codes each paragraph in each article in this dataset as containing either (1) no false or misleading claim or (2) one of 16 false or misleading claim labels outlined in the taxonomy of claims of the CARDS model in Appendix A.2. The generated claim labels combine the super-claim and sub-claim into a single label, delimited by an underscore ("_").

We sent requests to the fine-tuned model's chat completion endpoint using OpenAI Python library and formatted the requests in chat messages format containing the system and user messages as described in section 4.2. An example of the prompt structure is illustrated in Appendix A.7. A total of 856,722 paragraphs from 71,175 articles published by low credible domains between January and December 2022 were classified by the model.

After using the fine-tuned GPT-3.5-turbo model to classify claims on social media, we randomly selected a stratified sample of 914 paragraphs and their corresponding claim labels for manual evaluation by a pair of climate change communication experts. Half of the paragraphs were selected randomly from the generated claims that were labeled to have no claim and the other half was randomly selected using stratified sampling from all other types of claims in proportion to their distribution within the 2022 social media dataset.

We apply RoBERTa$_{large}$ CARDS model on the sampled paragraphs from social media by retrieving the model weights from Github[10] and classifying the claims in these paragraphs.

After classifying the claims in the sampled paragraphs from most engaging articles on social media using the fine-tuned GPT-3.5-turbo and RoBERTa$_{large}$ CARDS models, two climate communication experts manually coded for the claims in the same set of paragraphs. This provides a baseline to compare the efficacy of the aforementioned LLMs with respect to expert classification of climate misinformation. We elaborate on the coding procedure in the next section.

4.4 Expert annotation of claims in social media data

Two climate change communication experts, each with over 20 years of experience as professors and academic researchers on how climate science is communicated, annotated the sampled paragraphs from mostly engaging articles on social media (described in section 4.3) on the super-claim and sub-claim levels per the CARDS coding manual(Coan et al., 2021).

Each annotator separately reviewed each paragraphs and assigned a corresponding super-and sub-claims based on its content. Coding the claims on these two levels enable the annotators to detect claims at a more granular level, which allows for the identification and validation of super-claims based on the detected sub-claims. The annotators then consulted and resolved any disagreements to create a final reconciled dataset of expert labels for all 914 paragraphs. The resulting labels from the manual annotation process then were formatted to include the super-claim and sub-claim in a single label that resembles the formatting used to train CARDS and fine-tune the GPT-3.5-turbo model[11].

The resulting labels from annotating the sampled paragraphs by the climate communication experts established a baseline (i.e., ground truth) to evaluate the level of alignment between the claims classified by GPT-3.5-turbo and RoBERTa$_{large}$, and the expert annotation of these claims on the social media data.

In the results section, we (1) compare the performance of GPT models (GPT-4 and GPT-3.5-turbo) to the CARDS model to establish a performance baseline for classifying false or misleading claim, and (2) describe the level of alignment between the generated claims by each LLM and the expert annotations on the sampled paragraphs from the most engaging articles on social media that are published

[10] https://github.com/traviscoan/cards

[11] The formatting is similar to the one described in section 4.1 combining the super- and sub-claims into a single label delimited by '_'

by low credible sources.

5 Results

5.1 Evaluating the performance of GPT models on CARDS data

Using the zero-shot approach (described in section 4.1) to classify climate misinformation on CARDS test dataset, our findings indicate a comparable performance of GPT-4 to the CARDS model in classifying false or misleading claims about climate change (F1=0.74) as shown in Table 2. However, in comparison to CARDS, GPT-4 appears to have a higher rate of false positives (Precision=0.70) indicating that the model is being more conservative in classifying paragraphs with no claims as false or misleading claims about climate change compared to the CARDS models. In addition, GPT-4 had a higher recall than CARDS indicating the model's capability to correctly classify instances containing claims relevant to climate misinformation.

Though GPT-4 is a much larger model compared to RoBERTa$_{large}$, in terms of parameter size, it is less resource intensive for stakeholders in climate change discourse such as researchers, policymakers, or think tanks to utilize this model when classifying climate misinformation as it does not require substantial investment in computational resources or expertise to develop a customized model such as CARDS for classifying climate misinformation. However, detecting false or misleading claims does not include the models' capability to respond to or mitigate such claims unless they are augmented with relevant external knowledge as seen in use cases for fact-checking climate change claims(Leippold et al., 2024). Accordingly, future research is needed to explore and extend our work to determine whether models capable of detecting climate misinformation can augment general models, such as GPT-4, with information about which claims are false or misleading to help orient these general models toward guard-railing against such claims.

Although GPT-4 had comparable performance to CARDS in classifying false or misleading claims on the CARDS test dataset, fine-tuning GPT-3.5-turbo resulted in an even more performant model. The fine-tuned model (as described in section 4.2) has outperformed GPT-4 and RoBERTa$_{large}$ models on F1-score by 13.5% and 9.1%, respectively (see Table 2). Furthermore, the fine-tuned model has an uplift in precision compared to GPT-4 and RoBERTa$_{large}$ CARDS model by 25.7% and 7.3%, respectively. This improvement was at a cost of a slight decline in the recall of the fine-tuned model (81%) from GPT-4's 82%. Still, the fine-tuned model had a much better recall (81%) compared to RoBERTa$_{large}$ as shown in Table 2.

Metric	GPT-3.5-turbo	GPT-4	RoBERTa$_{large}$
Precision	0.88	0.70	0.82
Recall	0.81	0.82	0.75
F1-Score	0.84	0.74	0.77

Table 2: Comparing the performance of GPT-3.5-turbo and GPT-4 in classifying false or misleading claims about climate change in paragraphs belonging to the CARDS test set across three classification metrics: precision, recall, and F1-Score. The results were also evaluated against the RoBERTa$_{large}$ CARDS model on the same test set.

In the next section, we describe our findings from evaluating the alignment between the generated claim labels by GPT-3.5-turbo and RoBERTa$_{large}$) with respect to the annotated claims by two climate-communication experts on the sampled paragraphs from articles circulating on social media that are published by low credible sources.

5.2 LLM vs. Expert classification of claims

First we evaluated the intercoder-relability to ensure the alignment between the two experts by calculating Krippendorf's alpha. Coders scored $\alpha_{\text{Krippendorff}}$=.89 showing strong alignment between the two experts with respect to their coding of claims of the sampled paragraphs from articles about climate change circulating on social media. Then, using the claims generated by GPT-3.5-turbo and RoBERTa$_{large}$ we calculated the $\alpha_{\text{Krippendorff}}$ for each model with respect to expert annotations.

We found a much higher level of alignment ($\alpha_{\text{Krippendorff}}$=0.89) between the fine-tuned GPT-3.5-turbo and the climate research experts compared to RoBERTa$_{large}$ CARDS model ($\alpha_{\text{Krippendorff}}$=0.66). This suggests that the fine-tuned GPT-3.5-turbo can classify false or misleading claims about climate change with the approximate reliability as two senior climate change communication experts with over 20 years of experience. This opens up an opportunity for designing future AI systems that code and annotate climate misinformation at a scale with a human oversight.

Delving further into the performance comparison between GPT-3.5-turbo and the RoBERTa$_{large}$

Claim Code	Fine-tuned GPT-3.5-turbo				RoBERTa$_{large}$			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
0	0.94	0.94	0.94	491	0.76	0.96	0.85	491
1	0.71	1.00	0.83	24	0.71	0.62	0.67	24
2	0.93	0.93	0.93	14	1.00	0.79	0.88	14
3	0.57	1.00	0.73	8	0.57	0.50	0.53	8
4	0.95	0.94	0.95	274	0.93	0.63	0.75	274
5	0.99	0.89	0.94	103	0.90	0.62	0.74	103
Accuracy			0.94	914			0.81	914
Macro avg	0.85	0.95	0.88	914	0.81	0.69	0.74	914
Weighted avg	0.94	0.94	0.94	914	0.83	0.81	0.80	914

Table 1: Model performance comparison between the fine-tuned GPT-3.5-turbo and CARDS model in classifying false or misleading claims about climate change on a sample content from social media. Performance is measure assessed based on precision, recall, F1 scores. The claim labels corresponding to the claim codes are: (0) No claim, (1) global warming is not happening, (2) humans greenhouse gases are not causing global warming, (3) climate impacts are not bad, (4) climate solutions won't work, and (5) the climate movement and/or science are unreliable.

CARDS model with respect to expert annotations at the super-claim level of analysis, we find the macro averaged F1 score for the fine-tuned GPT-3.5-turbo (F1-macro=0.88) to be 18.9% higher than the one reported by RoBERTa$_{large}$ CARDS model (F1-macro=0.74) as shown in Table 1. The fine-tuned GPT-3.5-turbo also predominately had higher F1 scores across the five main categories of claims indicating strong performance by the model in identifying and classifying the main categories of claims outlined by the CARDS taxonomy, but on a broader sample of text from social media.

On the other hand, we observed a poor performance by the fine-tuned model in detecting claims related to climate impacts are not bad. Reviewing the annotated super-claims and sub-claims by experts, we found that the fine-tuned model is unable to accurately classify sub-claims within this category about the impacts of climate change on animal and plant species (see sub-claim 3.2 within the taxonomy of claims in Appendix A.2). We found that the fine-tuned GPT-3.5-turbo is biased toward inaccurately classifying claims regarding climate change impacts on animal and plant species as false, possibly due to biases in the CARDS training data that we originally fine-tuned the model on (see Limitations in section 7). This indicates additional fine-tuning is needed to enhance the ability of the model to differentiate between text describing positive versus negative impacts of climate change.

6 Conclusion

As developers and researchers test the potential of LLMs to persuade and misinform at scale (Matz et al., 2024; Zhang et al., 2024), evaluating the potential for LLMs to be part of the solution for online dis/misinformation rather than the problem becomes a task of great import. In this context, the overarching goal of this paper was to benchmark the ability of LLMs to classify climate change dis/misinformation that circulates on social media in comparison to other NLP computer-assisted approaches and climate change communication experts.

The results showed that though GPT-4 performance was inferior to trained BERT-based model in classifying climate change misinformation, a fine-tuned GPT-3.5-turbo model was superior to a trained BERT-based model and functionally equivalent as a climate change communication expert with 20 years of experience in classifying claims about climate change in social media. This demonstrates the potential for LLMs to be deployed in variety of governance roles to help mitigate potential harms that are inadvertently or purposefully perpetuated by LLMs, such as content moderation of social media or guardrailing general purpose LLMs from hallucinating and generating false information.

Our paper's results are also indicative of a broader shift in NLP approaches, training and fine-tuning LLMs for NLP tasks that previously relied on other computational approaches. Though LLMs are closed-source, employing fine-tuned LLM-based tools and applications may enhance the ability of civil society organizations that often have limitations on their technical expertise to engage in a range of important governance tasks (e.g., identifying and tracking dis/misinformation and hate speech).

7 Limitations

There are several limitations of this research. First, the data used for benchmarking GPT-4 and fine-tuning GPT-3.5-turbo is in English and in text format, which excludes claims in other languages and modalities (e.g., images and videos). This sets an important boundary condition on the performance of the models in identifying climate change misinformation while providing pathways for future research on LLM capabilities to accurately and reliably classify such content.

Another limitation is the reliance of our research on a single taxonomy of claims developed by the authors of CARDS (Coan et al., 2021) that was leveraged to fine-tune GPT-3.5.-turbo. This introduces two sets of biases. First, as the CARDS annotated dataset was based on an expert review of climate skeptic and contrarian domains, whether the GPT-3.5-turbo model is capable of precisely discriminating between accurate and inaccurate climate change claims within high credible sources (e.g. The New York Times, CNN, etc.) is an open question. One way to address this problem, as the authors of the CARDS model have recently moved toward, is a two-stage approach for classifying claims that first determines the veracity of the claim and in the second stage labels the category of false claims (Rojas et al., 2024). A next step, therefore, is to benchmark GPT-3.5-turbo model's performance in classifying claims from high credible sources for comparison to the updated CARDS model and continue fine-tuning as necessary on text sourced from domains with varying credibility.

A second model bias is the inability to classify claims in social media posts that do not fall within the CARDS taxonomy. For instance, the original taxonomy upon which CARDS is based includes claims about human health impacts as a category but was excluded in the final model due it its low prevalence (Coan et al., 2021). However, this may be a function of the ideological-skew of climate skeptic blogs on which CARDS was trained that ignored this dimension of climate change impacts and/or temporal trends increasing the prominence of health impacts. As a result, climate change communication experts annotating the social media claims observed a substantial number of false claims about the health impacts of climate change on humans that did not fall within the current CARDS model and which the GPT 3.5-Turbo was unable to classify.

In addition, the model's poor performance in classifying claims about the impacts of climate change on animal and plant species (see Table 1), could also be attributed to the under-representation of these examples in the CARDS dataset used for fine-tuning. Moving forward, fine-tuning the GPT-3.5-turbo model on additional expert annotated datasets, for example from Climate Feedback [12] would likely enhance the model's performance in accurately classifying a wider range of claims. These limitations stress the importance of benchmarking the performance of LLMs against data collected from "the wild" as we did in this paper and fine-tuning accordingly to ensure optimal performance in detecting misinformation online.

8 Ethics Statement

Incorporating the knowledge of domain experts into the design and development of AI tools for classifying the veracity of claims about climate change, or any other topic (e.g., politics, healthcare, or public policy), requires careful considerations of bias and impact. Frameworks or taxonomies of "truth" integrated with computer-assisted tools, regardless of their scientific basis, may have ideological or inadvertent subjective biases that narrow the range of information that is deemed accurate or inaccurate beyond what is optimal for free and open discourse. Therefore, it is important to mitigate such biases in the design and development process of AI-driven claim-detection and fact-checking tools by incorporating the inputs from diverse teams of researchers.

The deployment of AI tools, similar to the ones evaluated in this work, to detect false or misleading claims also have social implications. For instance, deploying these tools for content moderation of online platforms or for other governance tasks raises normative questions about free-speech that requires the engagement from a diverse range of societal stakeholders and decision-makers. It is also crucial to consider the potential exploitation and abuse of these by malicious actors, such as authoritarian regimes, to limit free expression. Mitigating this threat requires researchers and developers to be mindful of these considerations in the development of AI tools, actively engage with a diverse range of societal stakeholders in their development and deployment, and guard against their misuse by malign actors.

[12] https://science.feedback.org

References

Mohammad Majid Akhtar, Rahat Masood, Muhammad Ikram, and Salil S Kanhere. 2023. False information, bots and malicious campaigns: Demystifying elements of social media manipulations. *arXiv preprint arXiv:2308.12497*.

Joachim Allgaier. 2019. Science and environmental communication on youtube: Strategically distorted communications in online videos on climate change and climate engineering. *Frontiers in communication*, 4:446007.

CAAD. 2024. Underperforming & unprepared. Climate Action Against Disinformation's report highlights how platforms have responded to the EU legislation for online safety so far.

CCDH. 2021. The toxic ten: How 10 fringe publishers fuel 69% of digital climate change denial. Report.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120.

Tom Ellison and Brigitte Hugh. 2024. Climate security and misinformation: A baseline.

Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 us presidential election. *First Monday*.

Michael Fore, Simranjit Singh, Chaehong Lee, Amritanshu Pandey, Antonios Anastasopoulos, and Dimitrios Stamoulis. 2024. Unlearning climate misinformation in large language models. *arXiv preprint arXiv:2405.19563*.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Dimitrios Gounaridis and Joshua P Newell. 2024. The social anatomy of climate change denial in the united states. *Scientific Reports*, 14(1):2097.

IPCC. 2022. *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Summary for Policymakers. Cambridge University Press, Cambridge, UK and New York, USA.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Tanmay Laud, Daniel Spokoyny, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Climabench: A benchmark dataset for climate change text understanding in english. *arXiv e-prints*, pages arXiv–2301.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, et al. 2024. Automated fact-checking of climate change claims with large language models. *arXiv preprint arXiv:2401.12566*.

Stephan Lewandowsky. 2021. Climate change disinformation and how to combat it. *Annual Review of Public Health*, 42:1–21.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Thomas Marlow, Sean Miller, and J Timmons Roberts. 2020. Twitter discourses on climate change: exploring topics and the presence of bots.

SC Matz, JD Teeny, Sumer S Vaid, H Peters, GM Harari, and M Cerf. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

Sahal Shaji Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. 2023. Arabic mini-climategpt: A climate change and sustainability tailored arabic llm. *arXiv preprint arXiv:2312.09366*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Paulo Pirozelli, Marcos M. José, Igor Silveira, Flávio Nakasato, Sarajane M. Peres, Anarosa A. F. Brandão, Anna H. R. Costa, and Fabio G. Cozman. 2023. Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Preprint*, arXiv:2309.10945.

Cristian Rojas, Frank Algra-Maschio, Mark Andrejevic, Travis Coan, John Cook, and Yuan-Fang Li. 2024. Augmented cards: A machine learning approach to identifying triggers of climate change misinformation on twitter. *arXiv preprint arXiv:2404.15673*.

Ana Romero-Vicente. 2023. Platforms' policies on climate change misinformation. Factsheet.

Dietram A Scheufele. 2014. Science communication as political communication. *Proceedings of the National Academy of Sciences*, 111(supplement_4):13585–13592.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change.

Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. 2020. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5):e665.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Hong Tien Vu, Annalise Baines, and Nhung Nguyen. 2023. Fact-checking climate change: An analysis of claims and verification practices by fact-checkers in four countries. *Journalism & Mass Communication Quarterly*, 100(2):286–307.

Ziang Xiao, Wesley Hanwen Deng, Michelle S Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q Vera Liao. 2024. Human-centered evaluation and auditing of language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Kaicheng Yang and Filippo Menczer. 2024. Anatomy of an ai-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, 4.

Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1302–1305.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9717–9724.

A Appendix

A.1 Climate Change keywords

A full list of the compiled climate change keywords identified by climate experts that are used to scrape and filter relevant climate change articles: climate change - climate crisis - climate effects - climate hoax - climate policy - climate resilience - climate science - climate summit - global warming - greenhouse gas - greenhouse gases - IPCC - green energy - climate hypocrisy - paris agreement - paris climate - net zero - net-zero - COP26 - climate conversation - climate test - climate gap - climate activists - climate activist - clean energy - climate negotiations - climate deal - green new deal - climate conference - green technology - green tech - climate fearmongering - climate fears - climate anxiety - carbon capture

A.2 CARDS Taxonomy of Claims

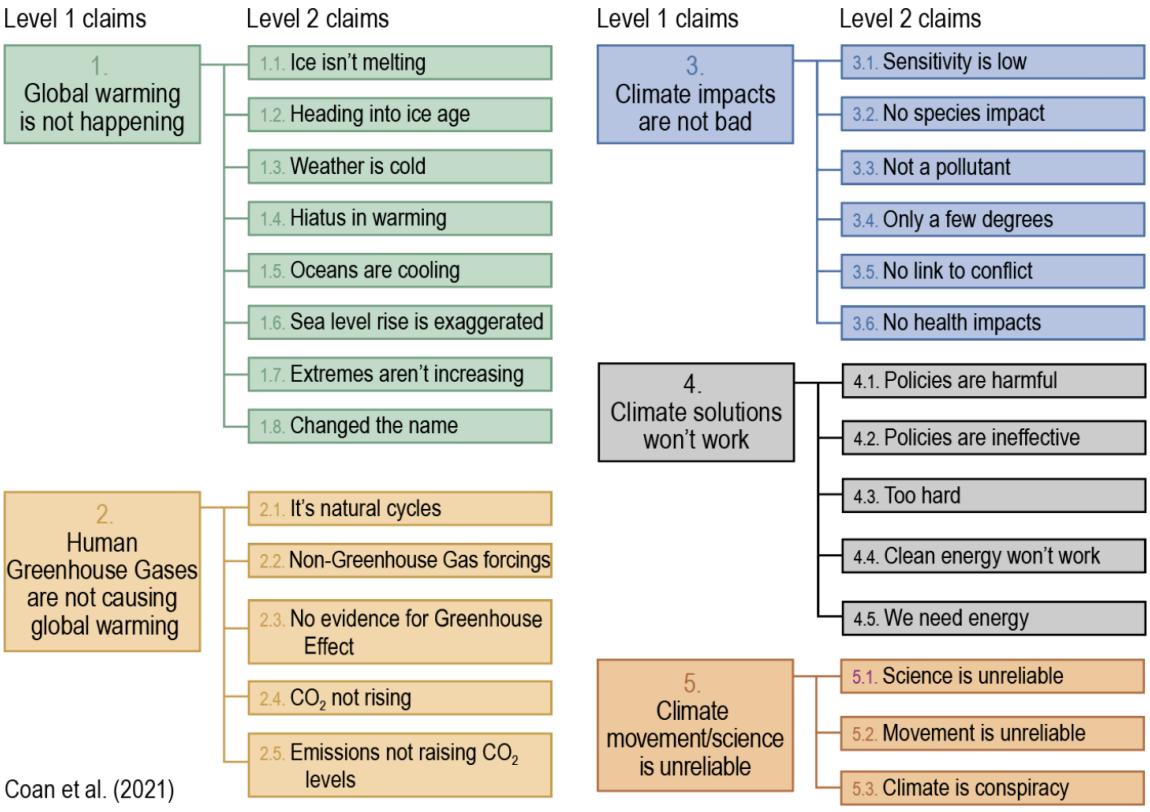

Figure 1: Taxonomy of claims published by (Coan et al., 2021)

A.3 Prevalence of domains in the social media data

Domain	Number of Articles	% Prevalence	Bias
newsbreak.com	208,855	24.37%	Questionable Source
freerepublic.com	47,236	5.51%	Right Bias
theepochtimes.com	32,536	3.79%	Questionable Source
foxnews.com	29,598	3.45%	Right Bias
beforeitsnews.com	25,824	3.01%	Questionable Source
breitbart.com	25,001	2.91%	Questionable Source
zerohedge.com	21,853	2.55%	Conspiracy-pseudocience
washingtonexaminer.com	18,348	2.14%	Right Bias
washingtontimes.com	15,614	1.82%	Questionable Source
patriotpost.us	14,488	1.69%	Right Bias
newsmax.com	11,571	1.35%	Questionable Source
americanthinker.com	10,500	1.22%	Questionable Source
wnd.com	10,053	1.17%	Questionable Sources
lawenforcementtoday.com	9,291	1.08%	Right Bias
shorenewsnetwork.com	9,016	1.05%	Right Bias
sott.net	8,560	0.99%	Conspiracy-pseudocience
dailycaller.com	8,479	0.98%	Right Bias
wattsupwiththat.com	8,202	0.95%	Conspiracy-pseudocience
bizpacreview.com	8,028	0.93%	Right Bias
townhall.com	7,910	0.92%	Questionable Source
lifesitenews.com	7,562	0.88%	Questionable Source
thelibertybeacon.com	7,464	0.87%	Conspiracy-pseudocience
noqreport.com	7,294	0.85%	Questionable source
dailywire.com	5,446	0.63%	Questionable Source
westernjournal.com	5,406	0.63%	Questionable Source

Table 3: Top 25 low credible domains, their prevalence, and bias category accounting for 65.85% of the total number of articles in the social media dataset described in Section 3.2.

A.4 System Prompt

```
"""
Overview:
---------
CARDS: Computer Assisted Recognition of Denial and Skepticism, is a machine learning project. Our aim is to train a computer to automatically detect and categorize misinformation about climate change. The end goal is that a computer can look at some text and successfully identify any climate misinformation - and even identify specific denialist claims. If successful, this will enable us to travel back in time and build a history of climate misinformation, including when myths originated and how they've evolved over time. It will also enable us to spot new publishing of denialist claims in real-time.

Context:
--------
Use the following coding rubric to answer the task assigned to you:
[
  {
    "code": "1",
    "identifier": 1,
    "claim": "Global warming is not happening"
  },
  {
    "code": "1_1",
    "identifier": 6,
    "claim": "Ice/permafrost/snow cover isn't melting"
  },
  {
    "code": "1_2",
    "identifier": 11,
    "claim": "We're heading into an ice age/ global cooling"
  },
  {
    "code": "1_3",
    "identifier": 12,
    "claim": "Weather is cold/snowing"
  },
  {
    "code": "1_4",
    "identifier": 13,
    "claim": "Climate hasn't warmed/changed over the last (few) decade(s)"
  },
  {
    "code": "1_5",
    "identifier": 14,
    "claim": "Oceans are cooling/not warming"
  },
  {
    "code": "1_6",
    "identifier": 15,
    "claim": "Sea level rise is exaggerated/not accelerating"
  },
  {
    "code": "1_7",
    "identifier": 16,
    "claim": "Extreme weather isn't increasing/ has happened before/isn't linked to climate change"
  },
  {
    "code": "1_8",
    "identifier": 17,
    "claim": "They changed the name from global warming' to climate change"
  },
  {
    "code": "2",
    "identifier": 2,
    "claim": "Human greenhouse gases are not causing climate change"
  },
  {
    "code": "2_1",
    "identifier": 18,
    "claim": "It's natural cycles/variation"
  },
  {
    "code": "2_2",
    "identifier": 24,
    "claim": "It's non-greenhouse gas human climate forcings (aerosols, land use)"
  },
  {
    "code": "2_3",
    "identifier": 25,
    "claim": "There's no evidence for greenhouse effect/carbon dioxide driving climate change"
  },
  {
    "code": "2_4",
    "identifier": 76,
    "claim": "C02 is not rising/ocean pH is not falling"
  },
  {
    "code": "2_5",
    "identifier": 78,
    "claim": "Human C02 emissions are miniscule/ not raising atmospheric C02"
  },
  {
    "code": "3",
    "identifier": 3,
    "claim": "Climate impacts/global warming is beneficial/not bad"
  },
  {
    "code": "3_1",
    "identifier": 31,
    "claim": "Climate sensitivity is low/negative feedbacks reduce warming"
  },
  {
    "code": "3_2",
    "identifier": 32,
    "claim": "Species/plants/reefs aren't showing climate impacts yet/are benefiting from climate"
  },
  {
    "code": "3_3",
    "identifier": 35,
    "claim": "C02 is beneficial/not a pollutant"
  },
  {
    "code": "3_4",
    "identifier": 37,
    "claim": "It's only a few degrees (or less)"
  },
  {
    "code": "3_5",
    "identifier": 38,
    "claim": "Climate change does not contribute to human conflict/threaten national security"
  },
  {
    "code": "3_6",
    "identifier": 39,
    "claim": "Climate change doesn't negatively impact health"
  },
  {
    "code": "4",
    "identifier": 4,
    "claim": "Climate solutions won't work"
  },
  {
    "code": "4_1",
    "identifier": 40,
    "claim": "Climate policies (mitigation or adaptation) are harmful"
  },
  {
    "code": "4_2",
    "identifier": 46,
    "claim": "Climate policies are ineffective/ flawed"
  },
  {
    "code": "4_3",
    "identifier": 53,
```

```
    "claim": "It's too hard to solve"
  },
  {
    "code": "4_4",
    "identifier": 55,
    "claim": "Clean energy technology/biofuels
    won't work"
  },
  {
    "code": "4_5",
    "identifier": 58,
    "claim": "People need energy (e.g., from
    fossil fuels/nuclear)"
  },
  {
    "code": "5",
    "identifier": 5,
    "claim": "Climate movement/science is
    unreliable"
  },
  {
    "code": "5_1",
    "identifier": 59,
    "claim": "Climate-related science is
    uncertain/unsound/unreliable (data, methods &
     models)"
  },
  {
    "code": "5_2",
    "identifier": 64,
    "claim": "Climate movement is alarmist/wrong/
    political/biased/hypocritical (people or
    groups)"
  }
]

Task:
-----
Classify whether a text excerpt belong to one of
     the claims outlined in the Context section
     provided in JSON format. Only respond in a
     JSON  format outlined below and don't make
     things up beyond what is given to you in the
     context. Below is the formatted JSON response
      template:
{
  "code": "CODE",
  "identifier": IDENTIFIER,
  "claim": "CLAIM"
}

If no claim is present in the text, just return a
      formatted json response like this one:
{
  "code": "0_0",
  "identifier": 0,
  "claim": "no claim"
}


This the end of the instructions. Now you will be
      provided a question with an excerpt of text
     and asked to identify the claim to which it
     belongs to.
"""
```

Listing 1: System prompt used for Zero-shot classification of claims on CARDS test data that is derived from CARDS coding manual in the supplimantary information of (Coan et al., 2021).

A.5 User Prompt

```
"""
Question: To what claim does the following text
     belongs to?
{text}

Answer:
"""
```

Listing 2: User prompt to generate the claim label corresponding to each paragrpah that is populated in the placeholder {text}.

```
"""
Question: To what claim does the following text
     belongs to?

What we are experiencing is outside of anything
     humans have seen on our planet and the only
     explanation that makes any real sense is that
      it is due to human actions.

Answer:
"""
```

Listing 3: An example user prompt illustrating how te paragraph is passed as part of the prompt

A.6 Prompts used for Fine-tuning

System prompt

```
"""
You are an expert in classifying false and
     misleading claims about climate change in
     news media. You are asked to classify whether
      a text excerpt belongs to one of the
     following labels separated by a comma: 0_0, 1
     _1, 1_2, 1_3, 1_4, 1_5, 1_6, 1_7, 1_8, 2_1, 2
     _2, 2_3, 2_4, 2_5, 3_1, 3_2, 3_3, 3_4, 3_5, 3
     _6, 4_1, 4_2, 4_3, 4_4, 4_5, 5_1, 5_2.
Your answer must only include the classification
     label with no additional details
"""
```

Listing 4: System prompt used as part of fine-tuning that describes for the model the task of classfying false or misleading claims about climate change.

Fine-tuning prompt structure

```
{
  "role": "system",
  "content": "You are an expert in classifying
     false and misleading claims about climate
     change in news media. You are asked to
     classify whether a text excerpt belongs to
     one of the following labels separated by a
     comma: 0_0, 1_1, 1_2, 1_3, 1_4, 1_5, 1_6, 1_7
     , 1_8, 2_1, 2_2, 2_3, 2_4, 2_5, 3_1, 3_2, 3_3
     , 3_4, 3_5, 3_6, 4_1, 4_2, 4_3, 4_4, 4_5, 5_1
     , 5_2. Your answer must only include the
     classification label with no additional
     details."
},
{
  "role": "user",
  "content": {text}
},
{
  "role": "assistant",
  "content": {claim}
}
```

Listing 5: A template request that includes the system, user, and assistant messags that were used to fine-tune the model. All requests were sent in JSON format that include all three messages.

A.7 Inference request using fine-tuned model

```
{
  "role": "system",
  "content": "You are an expert in classifying
     false and misleading claims about climate
     change in news media. You are asked to
     classify whether a text excerpt belongs to
     one of the following labels separated by a
     comma: 0_0, 1_1, 1_2, 1_3, 1_4, 1_5, 1_6, 1_7
     , 1_8, 2_1, 2_2, 2_3, 2_4, 2_5, 3_1, 3_2, 3_3
     , 3_4, 3_5, 3_6, 4_1, 4_2, 4_3, 4_4, 4_5, 5_1
     , 5_2. Your answer must only include the
     classification label with no additional
     details."
},
{
  "role": "user",
  "content": "What we are experiencing is outside
      of anything humans have seen on our planet
     and the only explanation that makes any real
     sense is that it is due to human actions"
},
{
  "role": "assistant",
  "content": ""
}
```

Listing 6: A sample request passed to the fine-tuned GPT-3.5-turbo model to generate the claim label corresponding to the paragraph in the user message.