

# Climate-Eval: A Comprehensive Benchmark for NLP Tasks Related to Climate Change

Murathan Kurfah<sup>†¶</sup> Shorouq Zahra<sup>†§¶</sup>  
 Joakim Nivre<sup>§¶</sup> Gabriele Messori<sup>§¶</sup>

<sup>†</sup>RISE Research Institutes of Sweden

<sup>§</sup>Uppsala University

<sup>¶</sup>Swedish Centre for Impacts of Climate Extremes (climes)

## Abstract

ClimateEval is a comprehensive benchmark designed to evaluate natural language processing models across a broad range of tasks related to climate change. ClimateEval aggregates existing datasets along with a newly developed news classification dataset, created specifically for this release. This results in a benchmark of 25 tasks based on 13 datasets, covering key aspects of climate discourse, including text classification, question answering, and information extraction. Our benchmark provides a standardized evaluation suite for systematically assessing the performance of large language models (LLMs) on these tasks. Additionally, we conduct an extensive evaluation of open-source LLMs (ranging from 2B to 70B parameters) in both zero-shot and few-shot settings, analyzing their strengths and limitations in the domain of climate change.

## 1 Introduction

Climate change represents one of the most pressing global challenges of our time, impacting every level of society, from international policy-making to everyday decisions. The importance of the topic is reflected in the vast amount of textual data generated, including a rich scientific literature as well as thousands of reports from corporations, government agencies and other organisations. Other data sources include countless social media posts and news articles capturing all aspects of the debate on climate change, from urgent calls for action to widespread misinformation. However, this enormous supply of unstructured information is not in a format amenable to analysis, and manual processing is unfeasible due to the sheer volume of text at hand.

A possible solution to this challenge comes from the field of Natural language processing (NLP). NLP has shown remarkable progress in recent years, notably with LLMs achieving near human

levels of performance across a variety of tasks. Thanks to their capacity to process textual data at scale, LLMs can help researchers process large volumes of data for a wide range of applications, such as analyzing climate-related documents or social media posts (El Barachi et al., 2021; Upadhyaya et al., 2023), structuring information about climate extremes from online texts into organized databases (Li et al., 2024; Madruga de Brito et al., 2025), and automatically detecting texts promoting climate change misinformation (Zhang et al., 2024; Zanartu et al., 2024). LLMs are thus playing a key role in enabling the compilation and analysis of climate change information from textual sources. In this context, it is essential to assess their performance. Such assessment needs to be comprehensive and consider a wide variety of tasks, since the performance of LLMs is known to be highly domain-dependent (Ling et al., 2023). In this paper, we aim to address this need by combining a newly-developed news classification dataset with existing datasets to create a unified benchmark that systematically evaluates the capabilities of LLMs across a vast array of climate-related NLP tasks. Our unified benchmark, ClimateEval, consists of 25 different tasks based on 13 datasets. It builds upon previous NLP benchmarking datasets, most notably ClimaBench (Spokoyny et al., 2023) and adds the following contributions:

- We introduce a new topic classification dataset using news articles on climate-related topics from the Guardian newspaper.
- We compile diverse climate-related NLP tasks into a unified benchmark using LM Evaluation Harness (Gao et al., 2024)<sup>1</sup>, facilitating systematic evaluation of models on various aspects of climate change discourse.

<sup>1</sup>The benchmark can be accessed here: <https://github.com/NLP-RISE/ClimateEval>

- We provide a comprehensive evaluation of a wide range of open-source LLMs on climate-related NLP tasks, highlighting these LLMs’ strengths and limitations as well as looking into the different challenges they pose.
- We provide a one-line evaluation setup to ensure accessibility and reproducibility for a wide range of users.

Through ClimateEval, we thus aim to facilitate NLP research by providing an easy-to-use setup enabling a comprehensive assessment of LLMs in climate change-related applications. By covering diverse, manually annotated datasets, our benchmark offers insights into the applicability of LLMs to critical aspects of climate discourse, from stance detection to claim verification.

## 2 Related Work

ClimaBench (Spokoyny et al., 2023) offers a collection of datasets designed to evaluate NLP models on climate-related tasks, such as text classification and stance detection. To the best of our knowledge, it is the first effort to aggregate multiple datasets into a benchmark for NLP models in the climate change domain, laying the foundation for ClimateEval. In addition to curating existing datasets, ClimaBench introduces a new dataset, CDP (Carbon Disclosure Project), which is based on climate-related questionnaires filled out by different stakeholders. ClimaBench was used for the evaluation of ClimateGPT (Thulke et al., 2024), an LLM specifically fine-tuned for climate-related applications. However, the ClimateGPT evaluation omitted or modified some of the ClimaBench tasks, including simplifying multi-class classification tasks into binary ones. Nonetheless, ClimateGPT incorporates two additional datasets not included in ClimaBench in its evaluation suite, PIRA (Pirozelli et al., 2024) and Exeter (Coan et al., 2021), which we also include in our benchmark. Fore et al. (2024) rely on question answering (QA) datasets to detect climate change misinformation in LLMs, specifically LLMs that have been intentionally injected with false climate information and subsequently made to unlearn it.

The above work provides a context for evaluating LLMs on climate-related tasks. A key knowledge gap is that only a limited number of models have been benchmarked, and some tasks have suffered reductions (e.g., transforming multi-

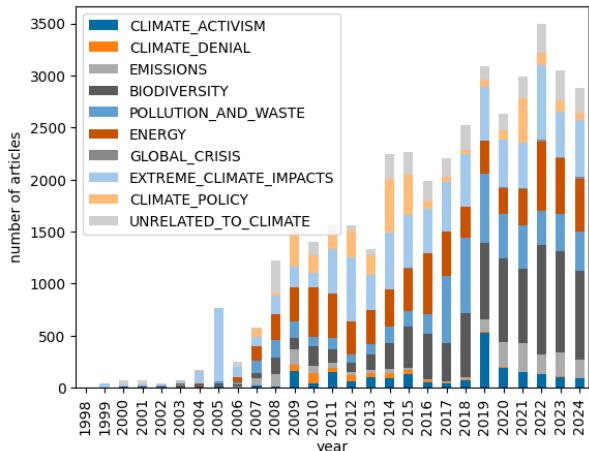


Figure 1: The distribution of articles by year in the Guardian Climate News Corpus, broken down by category.

class classification into a binary classification problem) without considering the effect on performance when transforming the task and target output in this fashion. ClimateEval complements and extends previous efforts by addressing this gap. Furthermore, the unified benchmark provides in unprecedented breadth of both tasks and annotated data.

## 3 The ClimateEval Benchmark

In this section, we detail the datasets and the corresponding tasks included in the ClimateEval benchmark, along with the evaluation metrics we employ.

### 3.1 Datasets and Tasks

In this subsection, we first describe our new dataset, the Guardian Climate News Corpus, followed by the other datasets included in ClimateEval.

- **Guardian Climate News Corpus :** A dataset containing climate-related and non-climate-related articles. These are assigned to nine climate-related categories (*climate activism*, *climate denial*, *emissions*, *biodiversity*, *pollution and waste*, *global crisis*, *extreme climate impacts*, and *climate policy*) and one *unrelated to climate* category with articles sampled from diverse domains (e.g., sports, technology, gardening).<sup>2</sup>

Data for each category is scraped from the Guardian’s website<sup>3</sup>. The Guardian was cho-

<sup>2</sup>The benchmark and full details on the categories and tags can be found on HuggingFace: [https://huggingface.co/datasets/NLP-RISE/guardian\\_climate\\_news\\_corpus](https://huggingface.co/datasets/NLP-RISE/guardian_climate_news_corpus).

<sup>3</sup>Using the Guardian Open Platform API <https://open-platform.theguardian.com/documentation/>

Dataset (Source)	Task(s)	# labels	Train	Dev	Test
ClimaText (Varini et al., 2021)	Sentence classification	2	121847	3918	5426
Climate-Stance (Vaid et al., 2022)	Stance classification	3	2871	354	355
Climate-Eng (Vaid et al., 2022)	Topic classification	5	2871	354	355
Climate-FEVER (Diggelmann et al., 2020)	Claim verification	3	-	-	7675
SciDCC (Mishra and Mittal, 2021)	Topic classification by Title	20	9231	1154	1154
	Topic classification by Title & Summary	20	9231	1154	1154
	Topic classification by Title & Body	20	9231	1154	1154
CLIMA-CDP (Spokoyny et al., 2023)	QA-Cities (answer relevance)	2	288418	51018	55872
	QA-Corp. (answer relevance)	2	207450	22044	29892
	QA-States (answer relevance)	2	52287	5814	6888
	Topic-Cities (topic classification)	12	46803	8771	8984
PIRA 2.0 MCQ (Pirozelli et al., 2024)	PIRA with Context	5	1798	225	227
	PIRA without Context	5	1798	225	227
Exeter Misinformation (Coan et al., 2021)	Claim Detection	6	23436	2605	2904
	Sub-claim Detection	18	23436	2605	2904
Climate-Change NER (Bhattacharjee et al., 2024)	Entity recognition	13	31633	6366	5775
CheapTalk (Bingler et al., 2023)	Climate Detection	2	1300	-	400
	Climate Sentiment	3	1000	-	320
	Climate Commitment	2	1300	-	400
	Climate Specificity	2	1000	-	320
	TCFD Recommendations	5	1300	-	400
Net-Zero Reduction (Schimanski et al., 2023)	Paragraph Classification	3	3441	-	-
Environmental Claims (Stammbach et al., 2023)	Sentence Classification	2	2400	300	300
Guardian Climate News Corpus	Topic classification by Title	10	32138	4017	4018
	Topic classification by Body	10	32138	4017	4018

Table 1: Overview of the ClimateEval benchmark tasks, subtasks, and dataset sizes. The **# labels** column gives the number of labels per task; for PIRA, it represents the number of answer choices per question, while for Climate-Change-NER, it corresponds to the total number of distinct entities annotated in the dataset.

sen because it explicitly permits the use of its content for research and non-commercial purposes. Accordingly, we are able to freely generate and distribute this dataset. Articles are scraped based on having been assigned specific tags. These tags are used by the Guardian to taxonomize their own publications; each tag is usually comprised of a section and unique topic identifier, separated by a forward slash (e.g., the “environment/flooding” tag is assigned to articles covering flooding-related incidents in the Guardian’s Environment section<sup>4</sup>). This taxonomy has enabled us to curate a manual selection of tags that are relevant to each of the ten categories with ease.

As example, an article that we categorize as falling under *climate activism* is scraped based on a list of article tags that relate to climate activism (such as “environment/school-climate-strikes”). However, this article could also have been assigned other tags that describe it, such as “australia-news/australian-education” and “world/extreme-weather” – the latter is

a tag that falls under our “Extreme Climate Impacts” category. Since the second tag fall under another category we have defined, we remove this article from the pool so that each article can only belong to one category; otherwise, no action is taken. After these are filtered out, any articles with a body shorter than 49 words or longer than 1,000 words are also dropped from the dataset, resulting in 40,173 articles in ten mutually-exclusive categories published any time between 1998 and 2024. Figure 1 shows the distribution of articles across years.

We derive two tasks from this dataset:

- i. *Topic classification by Title*, where multi-class classification is performed only on the title of the article; and
- ii. *Topic classification by Body*, where the same task is performed on the body of the article.

We believe the Guardian Climate News Corpus is a valuable addition to existing resources, as it provides a large-scale, real-world news

<sup>4</sup><https://www.theguardian.com/uk/environment>

dataset focused on climate topics with fine-grained labels that capture diverse aspects within climate discourse. In contrast, existing datasets, such as SciDCC (Mishra and Mittal, 2021), are not exclusively focused on climate change, but instead cover broader scientific categories, of which some are climate-related (e.g., “Pollution” and “Hurricanes”).

For simplicity, we have chosen to create a dataset to evaluate multi-class classification as each article belongs only to a single class. However, it is possible to reproduce this dataset with a different set of tags, with the option to not filter out articles sharing tags across categories, in turn creating a dataset more suitable for multi-label classification.<sup>5</sup>

- **ClimaText** (Varini et al., 2021): A dataset with sentences from the web, Wikipedia, and public companies’ 10-K reports. Each sentence is labeled for whether it is related to climate change or not. Therefore, this dataset is suitable for the task of *binary classification of sentences*, requiring models to distinguish texts relevant to climate change.
- **Climate-Stance** (Vaid et al., 2022): A dataset consisting of 3,777 tweets posted during the 2019 United Nations Framework Convention on Climate Change COP (Conference of the Parties). The dataset is annotated for stance detection (classification), where each tweet is categorized into one of three classes: (i) being in *favor* of climate change mitigation, (ii) being *against* such measures, or (iii) taking an *ambiguous* stance.
- **Climate-Eng** (Vaid et al., 2022): A dataset designed for multi-class topic classification, constructed from the same set of 3,777 tweets as the Climate-Stance dataset, but labeled according to one of five distinct topics: *disaster*, *ocean/water*, *agriculture/forestry*, *politics*, or *general*.
- **Climate-FEVER** (Diggelmann et al., 2020): A claim verification dataset that consists of real-world claims about climate change. Each of the 1,535 claims is paired with five evidence sentences extracted from Wikipedia,

<sup>5</sup>The code used for scraping the dataset from the Guardian’s Open Platform is available on GitHub: <https://github.com/NLP-RISE/extractguardian>

which either *support*, *refute*, or *provide insufficient information* about the claim. Climate-FEVER also includes a general label, determined by aggregating the ratings of individual claim-evidence pairs. However, in our benchmark, we model the task for this dataset as a three-way entailment problem where each claim is evaluated individually against each of its five evidence sentences, resulting in 7,675 claim-evidence pairs.

- **SciDCC** (Mishra and Mittal, 2021): A multi-class classification dataset comprising news articles sourced from the Science Daily website, annotated with one of the possible 20 scientific categories (e.g., *biology*, *weather*, *ozone holes*, *endangered animals*). Each article includes a title, a summary, and a body section. In order to fully utilize the available information at different granularity, we model the task in three ways:
  - i. *Topic Classification by Title*, where classification is performed solely on the article title;
  - ii. *Topic Classification by Title & Summary*, which utilizes both the title and the summary sections; and
  - iii. *Topic Classification by Title & Body*, where the full article body is used, in addition to the title.
- **CLIMA-CDP** (Spokoyny et al., 2023): A dataset derived from disclosure questionnaires collected and made available by Carbon Disclosure Project (CDP), an international non-profit organization. These questionnaires are completed by various stakeholders, including cities, corporations, and states. The dataset consists of responses to hundreds of unique questions related to climate impact, mitigation efforts and governance, and supports two distinct classification tasks:
  - i. *CDP-QA* is a binary classification task that predicts whether a given report response correctly answers the questions posed. This task has three variants based on the type of stakeholder providing

*CDP-QA* is a binary classification task that predicts whether a given report response correctly answers the questions posed. This task has three variants based on the type of stakeholder providing

- the response: *CDP-QA-Cities*, *CDP-QA-Corporations*, and *CDP-QA-States*, for municipal, corporate and state-level responses, respectively.
- ii. *CDP-Topic-Cities* is a multi-class classification task where responses from city stakeholders are categorized into one of twelve predefined topics (e.g., *climate hazards*, *emissions*, *energy*, *food*). The task is limited to city responses since no annotations exist for other stakeholders.
- **PIRA 2.0 MCQ** ([Pirozelli et al., 2024](#)): A multiple-choice QA dataset constructed from a collection of scientific abstracts and United Nations reports, with a focus on climate-related topics such as oceanography, coastal ecosystems, and climate change impacts. Each instance consists of a question and five answer choices. Additionally, each question is accompanied by a supporting context that provides relevant information to guide the answer selection process. We divided the dataset into two sub-tasks:
    - i. *PIRA with Context*, where models are expected to answer a question with the help of the accompanying context, like in an open-book exam setting;
    - ii. *PIRA without Context*, where models are required to answer the question without any supporting information.
- By structuring the dataset in this manner, we aim to evaluate both retrieval-augmented and self-knowledge-based approaches to climate-related QA.
- **Exeter Misinformation** ([Coan et al., 2021](#)): A dataset designed to detect climate misinformation by annotating text from prominent climate contrarian blogs and think tanks spanning over 20 years (1998–2020). The dataset is structured according to a two-level taxonomy of climate contrarian claims. The first level consists of broad claim categories, such as: (1) *Global warming is not happening*, which is further divided into more specific sub-claims, including (1.1) *Ice isn't melting*, or (1.2) *Oceans are cooling*. To test different levels of misinformation detection, we define two sub-tasks:
    - i. *Claim Detection*, a classification task where texts are categorized into one of the six first-level claim labels; and
    - ii. *Sub-Claim Detection*, a more fine-grained classification task where each text is assigned to its corresponding sub-claim category within the taxonomy.
- These tasks allow for a detailed evaluation of how LLMs classify misinformation narratives in climate discourse at different granularities.
- **Climate-Change NER** ([Bhattacharjee et al., 2024](#)): A *named entity recognition* (NER) dataset constructed from 534 scientific abstracts sourced from the Semantic Scholar Academic Graph ([Kinney et al., 2023](#)). The dataset was collected using a set of climate-related keywords (e.g., *wildfire*, *floods*) to ensure relevance to climate science. Each abstract is annotated for entity types that are specific to the climate discourse, such as *greenhouse gases*, *climate hazards*, *climate-impacts*, *climate mitigations*. The associated task is a token classification task, where the goal is to identify tokens corresponding to these entities. NER is a crucial preprocessing step for various downstream applications that aim to extract structured information from unstructured text ([Li et al., 2024](#)). Hence, a model's performance on climate-specific NER is a strong indicator of its usefulness for information extraction in climate discourse ([Li et al., 2020](#)).
  - **CheapTalk** ([Bingler et al., 2023](#)): A dataset consisting of corporate disclosures, focusing on how companies communicate climate-related information. Unlike the previously described datasets, primarily dealing with either academic discourse or social media posts, this dataset uncovers corporations' climate communication strategies. The following five tasks are performed on it:
    - i. *Climate Detection*: A binary classification task to determine whether a given text passage discusses climate-related topics. This dataset serves as a filtering mechanism to extract climate-relevant content from corporate reports.
    - ii. *Climate Sentiment*: A sentiment analysis task that categorizes climate-related

Task	Gemma-2		Llama-2		Llama-2		Climate GPT-7B		Climate GPT-13B		Qwen-2.5		Llama-3.1		Mistral 24B		Llama-3.3	
	2B	7B	7B	13B	0	5	0	5	0	5	0	5	0	5	0	5	0	5
CDP-QA-Cities	.46	.60	.62	<b>.69</b>	.37	.55	.48	.57	.59	.57	.61	.62	.57	.66	<b>.68</b>	.68	.61	<b>.69</b>
CDP-QA-Corp.	.46	.59	.62	<u>.67</u>	.32	.55	.40	.57	.58	.58	.60	.61	.51	.64	<u>.67</u>	<b>.68</b>	.60	<b>.68</b>
CDP-QA-States	.46	.61	.68	<u>.71</u>	.37	.56	.48	.59	.60	.58	.62	.63	.54	.66	<b>.72</b>	.68	.59	.70
CDP-Topic-Cities	.28	.34	.32	<u>.35</u>	<u>.35</u>	<u>.35</u>	.33	<u>.35</u>	.30	<u>.35</u>	.30	<u>.35</u>	.31	<b>.37</b>	.34	.32	<u>.35</u>	.35
Climate Commit.	.42	.63	.66	<b>.71</b>	.56	.63	.63	.65	.61	.60	.62	.63	.54	.66	<b>.71</b>	<b>.71</b>	.68	.68
Climate Detection	.20	.70	.65	.69	.61	.60	.59	.66	.62	.68	.53	.73	.50	.76	.69	<b>.78</b>	.71	.76
Climate Eng	.37	.46	.52	<u>.59</u>	.37	.50	.52	.53	.53	.50	.52	.54	.55	<u>.59</u>	<b>.60</b>	.59	.57	.59
Climate NER	.11	.19	.14	.20	.10	.17	.09	.17	.07	.14	.04	.18	.16	.21	.23	<b>.30</b>	.20	.27
Climate Sentiment	.49	.64	.72	<u>.74</u>	.44	.71	.60	.70	.61	.65	.57	.63	.54	.72	<b>.75</b>	.72	.73	<u>.75</u>
Climate Specificity	.39	.66	.72	.69	.54	.60	.51	.64	.48	.60	.53	.68	.60	.73	<u>.75</u>	<b>.79</b>	.75	<b>.79</b>
Climate Stance	.13	.50	.30	.54	.20	.47	.16	.48	.11	.37	.14	.42	.06	.56	.18	.61	.26	.64
Climate-Fever	.33	.34	<b>.57</b>	.48	.40	.34	.35	.41	.25	.45	.41	.55	.52	.51	.56	.50	.51	.55
Climatext Sent Clf.	.39	.64	.56	.62	.58	.67	.63	.70	.58	.61	.54	.68	.47	.68	.57	.68	.63	<b>.71</b>
Env. Claims	.53	.77	.80	.81	.61	.79	.66	.78	.80	.75	.64	.81	.75	<u>.83</u>	.75	.82	<b>.85</b>	.83
Exeter Claim	.14	.41	.43	.48	.17	.30	.20	.38	.26	.37	.34	.41	.35	.46	.55	<b>.61</b>	.56	.59
Exeter Sub-Claim	.22	.27	.48	.51	.05	.13	.15	.25	.24	.25	.33	.40	.39	.47	<u>.61</u>	<b>.63</b>	.59	.59
Guardian Body	.39	.31	.48	<u>.54</u>	.29	.19	.31	.21	.30	.25	.33	.45	.47	.49	<b>.58</b>	.00	.00	.00
Guardian Title	.35	.35	.45	<u>.54</u>	.40	.41	.38	.00	.46	.48	.45	.00	.42	.50	.51	.53	<b>.57</b>	.00
Net-Zero Reduction	.22	.44	.38	.45	.46	.84	.39	<u>.86</u>	.23	.83	.43	.85	.48	.43	.28	.47	.60	<b>.89</b>
Pira W/Ctx.	.86	.86	<u>.95</u>	.94	.74	.70	.87	.88	.86	.87	.93	.94	.93	<u>.95</u>	.94	<b>.96</b>	.94	.92
Pira W/O Ctx.	.64	.62	.84	.84	.44	.52	.58	.67	.67	.74	.70	.81	.69	.80	<b>.89</b>	<b>.89</b>	.88	.88
SciDCC Title	.16	.19	.27	<u>.28</u>	.18	.23	.08	.20	.16	.17	.16	.24	.15	.23	<u>.28</u>	<b>.33</b>	.25	<u>.33</u>
SciDCC Title Body	.10	.21	.20	.23	.13	.14	.08	.17	.13	.11	.17	.18	.18	.23	.26	<b>.32</b>	.26	.31
SciDCC Title Sum.	.10	.19	.25	.27	.13	.25	.12	.25	.17	.16	.20	.25	.20	.27	.28	<u>.33</u>	.27	<u>.34</u>
Tcfdf Recommend.	.15	.31	.43	.46	.24	.30	.24	.30	.16	.29	.29	.37	.26	.45	.50	<u>.52</u>	<b>.53</b>	.48

Table 2: Evaluation results for various models across different few-shot experiments. The numbers indicate the models’ performance for each task in F1-macro, except for PIRA (see Section 3.2). Numbers in boldface are the highest performing, whereas underlined numbers are the second highest for that task. For shortened task identifiers, see Table 6.

statements based on their tone. Each passage is labeled as highlighting risks, emphasizing opportunities, or maintaining a neutral stance on climate change.

- iii. *Climate Commitments*: A classification task that determines whether a corporate disclosure contains a climate-related commitment. Texts are labeled as *commitment-yes* if they explicitly state planned or ongoing climate actions and *commitment-no* if they do not reference concrete climate actions.
- iv. *Climate Specificity*: A binary classification task assessing the specificity of corporate climate commitments. A passage is labeled as *specific* if containing “detailed performance information, details of actions, or tangible and verifiable targets” (Bingler et al., 2023), or *non-specific* if it is vague.
- v. *TCFD Recommendations*: A multi-class classification task assessing corporate disclosures against the guidelines of the Task Force on Climate-related Financial Disclosures (TCFD), an international framework that standardizes the reporting of climate-related financial risks and

opportunities.<sup>6</sup> Each text is labelled by one of the four TCFD recommendation categories (*governance*, *strategy*, *risk management*, and *metrics and targets*) or as *none* if none applies.

- **Net-Zero Reduction** (Schimanski et al., 2023): A dataset comprising 3,517 expert-annotated paragraph samples designed to detect and assess net-zero and emission reduction targets in corporate, national, and regional communications. Each sample is labeled as *net-zero target* (commitment to net-zero emissions), *reduction target* (commitment to emission reduction without full net-zero), or *no target* (no explicit reduction commitment). This dataset is used for multi-class classification.
- **Environmental Claims** (Stammbach et al., 2023): A dataset comprising 3,000 sentences from corporate sustainability reports, earnings calls, and annual reports by publicly listed companies that are expert-annotated for environmental claims. An environmental claim any statement suggesting that a product, service, or company is environmentally friendly. Each sentence is labeled as containing an en-

<sup>6</sup><https://www.fsb-tcfd.org>

vironmental claim or not. This dataset is used for a binary sentence classification task.

Most of the datasets in ClimateEval were developed in the pre-LLM era and were not designed with prompting-based evaluation in mind. Moreover, prior work has used some datasets in different ways — for example, ClimateGPT modeled Climate-FEVER as a binary classification task (not three-way). To address this lack of standardization, ClimateEval provides a unified evaluation benchmark that standardizes task formulations, label sets, and prompts. Each task is paired with a suitable prompt and, where applicable, modeled at multiple granularities (e.g., title-only vs. full-body classification in SciDCC). This unification ensures that different models are evaluated under consistent conditions, enabling reproducible comparisons across tasks.

### 3.2 Evaluation Setup and Metrics

ClimateEval is implemented using the LM Evaluation Harness library (Gao et al., 2024), which provides an easy-to-use infrastructure for evaluating language models on a wide range of tasks. Each task in the benchmark is defined through a YAML-formatted configuration file, specifying the input format, prompt template, expected output and the target metric. The benchmark can be executed with a single command, allowing for efficient and standardized evaluation across diverse models and tasks.

We evaluate model performance primarily using accuracy and macro-averaged F1-score for all classification tasks. Macro F1 is particularly important due to class imbalances in several datasets where there are as many as 20 labels with skewed distributions (see Appendix C for details). Macro F1 averages scores across all classes equally to ensure a balanced assessment by preventing frequent labels from biasing the results.

For Climate-Change-NER, the only sequence labeling task in our benchmark, we compute precision, recall, and F1-score based on the entity-type and entity-span pairs. Each entity-type is evaluated independently, and a model’s prediction is considered correct if it correctly identifies an entity within the set of gold entities for that type. If a gold-standard entity-type contains multiple entities and the model predicts only a subset, we count each correctly identified entity as a true positive, while missing entities contribute to false negatives. We

report only exact matches, where an entity is correct only if both the span and type match perfectly.

For multiple-choice QA (MCQA) tasks (e.g., PIRA 2.0 MCQ), we report exact match accuracy, as F1-score is not meaningful in this context. In MCQA tasks, each question has one correct answer, and the model selects from arbitrary option labels (e.g., A, B, C), which are not semantic classes. Since the model selects a single arbitrary option and only one answer is correct, precision and recall are not meaningful, making exact match accuracy the appropriate metric.

## 4 Evaluation of Open-Source LLMs

We report the performance of a range of open source baseline LLMs with varying sizes, ranging from 2B to 70B, in both zero-shot and 5-shot scenarios. The mid-sized models (Mistral 24B and Llama3.3-70B) are loaded in 4-bit quantization whereas the other models were run in half-precision FP16.

For classification tasks, the log-likelihoods of each possible label are calculated, and the label with the highest likelihood is selected as the model’s prediction. For the generation task, Climate Change NER, the model is simply prompted to generate the corresponding JSON file.

The baseline LLMs that we use are: Gemma-2 (2B) (Team et al., 2024), Qwen-2.5 (7B) (Yang et al., 2025), ClimateGPT (both 7B and 13B) (Thulke et al., 2024); Llama-2 (both 7B and 13B); Llama-3.1 (8B), Llama-3.3 (70B) (Grattafiori et al., 2024) and Mistral (24B). The baseline models’ performances are reported in Table 2 in both zero-shot and five-shot settings.

### 4.1 Zero-shot vs. Few-shot Performance

Across the benchmark, few-shot prompting consistently improves performance (Figure 2). The most significant gains are observed in Climate Stance (+0.26 for ClimateGPT-7B to +0.50 for Llama-3.1-8B) and Net-Zero Reduction (+0.15 for Qwen-2.5-7B to +0.44 for Gemma-2-2B). Both tasks rely on understanding specialized classification taxonomies, whether for categorizing social media discourse or analyzing policy documents. The improvement diminish as the number of labels increases. The Exeter Sub-claim Detection and SciDCC tasks exhibit minimal gains (0.03 on average), suggesting that when models must choose between a high number of categories, few-shot

Task	0-Shot	5-Shot	Diff.
CDP-QA-Cities	.55 (.10)	.63 (.06)	.07
CDP-QA-Corp.	.53 (.11)	.62 (.05)	.09
CDP-QA-States	.56 (.11)	.64 (.06)	.07
CDP-Topic-Cities	.32 (.02)	.35 (.01)	.03
Climate Commit.	.61 (.09)	.67 (.04)	.06
Climate Detection	.57 (.15)	.71 (.06)	.14
Climate Eng	.50 (.08)	.54 (.05)	.04
Climate NER	.13 (.06)	.20 (.05)	.08
Climate Sentiment	.60 (.11)	.70 (.04)	.09
Climate Specificity	.58 (.13)	.69 (.07)	.10
Climate Stance	.17 (.07)	.51 (.09)	.34
Climate-Fever	.43 (.11)	.46 (.08)	.03
ClimaText Sent Clf.	.55 (.08)	.67 (.03)	.12
Env. Claims	.71 (.10)	.80 (.03)	.09
Exeter Claim	.33 (.16)	.45 (.10)	.11
Exeter Sub-Claim	.34 (.19)	.39 (.17)	.05
Guardian Body	.42 (.13)	.41 (.18)	-.01
Guardian Title	.44 (.07)	.48 (.08)	.04
Net-Zero Reduction	.39 (.13)	.67 (.22)	.29
Pira W/ Ctx.	.89 (.07)	.89 (.08)	.00
Pira W/O Ctx.	.70 (.14)	.75 (.13)	.06
SciDCC Title Sum.	.19 (.07)	.26 (.06)	.07
SciDCC Title	.19 (.07)	.24 (.06)	.06
SciDCC Title Body	.17 (.06)	.21 (.07)	.04
TCFD Recommend.	.31 (.14)	.39 (.09)	.07
Average	.45	.53	.09

Table 3: Comparison of average model performance across tasks in 0-shot and 5-shot settings. Standard deviations are provided within parentheses. The last column shows the performance difference between the zero-shot and five-shot experiments. For shortened task identifiers, see Table 6.

prompting does not provide sufficient guidance.

Another set of tasks that do not benefit from in-context learning includes Climate-FEVER (fact verification) and PIRA (multiple-choice QA), which show no improvement. Unlike standard classification tasks, these tasks require external knowledge to be successfully carried out, limiting the efficiency of few-shot prompting. As each instance in these datasets depends on specific knowledge, the lack of improvement is unsurprising.

## 4.2 Analysis of tasks

Among the 25 tasks in our benchmark, certain tasks consistently emerge as more challenging than others (Table 3). SciDCC and Climate-Change-NER exhibit particularly poor performance, with F1-scores below 0.34 across models, even with few-shot prompting. The poor performance on SciDCC can be attributed to the large number of classes (20 in total), making the assignment more difficult. Surprisingly enough, classification of news texts based solely on their titles only leads to a minor performance loss compared to using the full article.

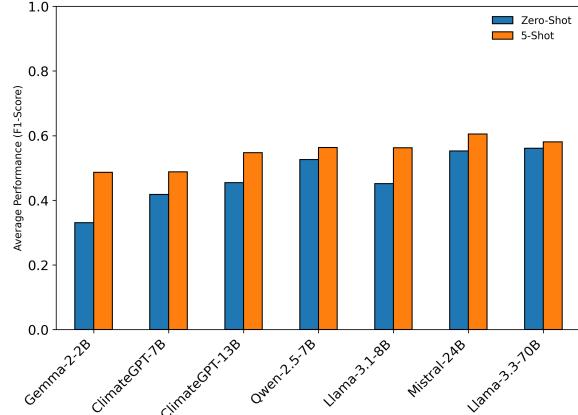


Figure 2: Average model performance across all tasks in Zero-Shot and 5-Shot settings.

This may indicate that titles provide sufficient cues regarding the content of articles.

Climate-Change NER, on the other hand, is the only non-classification task in the benchmark, requiring models to correctly identify climate-related named entities in scientific abstracts. This task involves both span detection and entity type classification, where entity types are domain-specific<sup>7</sup> and differ significantly from standard NER labels such as "location" or "organization". Without specific training/fine-tuning on climate-specific entity annotations, it is understandable that general-purpose LLMs struggle to accurately extract these entities.

In contrast, the LLMs perform strongly on several tasks in the benchmark, even in the zero-shot setting. The performance on PIRA improves by 0.2 points on average in the zero-shot setting when relevant context is provided, showcasing the LLMs' ability to identify and utilize relevant parts of additional knowledge for question answering (QA). However, the overall high performance on PIRA even without any context suggests that the task may be relatively easy, highlighting the need for a more challenging QA dataset.

Perhaps the most surprising finding is the relatively poor performance of the models on the CheapTalk dataset tasks. Intuitively, one might expect tasks like Climate Detection (whether or not a text is climate-related) or Climate Sentiment (classifying tone as risk, opportunity or neutral) to be straightforward for advanced LLMs. Yet, our results show that even the best models achieve F1 scores below 0.8 on these tasks, indicating that

<sup>7</sup>For the full list of entities, see: <https://huggingface.co/datasets/ibm-research/Climate-Change-NER>

these models experience challenges in understanding the climate discourse.

### 4.3 Impact of In-domain Training

ClimateGPT presents a unique opportunity to evaluate the impact of in-domain training for climate-related NLP, as it is an LLM developed through continuous pre-training of Llama-2 models on 4.2B climate-focused tokens and instruction-tuned on expert-curated datasets (Thulke et al., 2024). To gain insights regarding the benefits of domain adaptation, we compare ClimateGPT against its base Llama-2 models (Table 4).

Table 3 shows the performance differences between these models. On average, ClimateGPT exhibits a slight improvement over Llama-2 by  $\leq 0.05$ . The largest gains are seen in Exeter Sub-claim, CDP-QA, and PIRA With Context. However, the improvements are inconsistent, with some tasks, including Climate-Change NER, Climate Stance, and Net-Zero Reduction, showing negligible or negative differences. The limited impact of in-domain training across tasks suggests that, while climate-focused continuous pre-training provides benefits in specific cases, the performance gains are not uniform across the tasks in ClimateEval.

## 5 Conclusion

We have presented ClimateEval, a comprehensive benchmark for climate change NLP, encompassing 13 datasets and 25 tasks that cover a wide range of climate-related language understanding tasks. By unifying these diverse tasks into a single framework, ClimateEval enables systematic assessment of how well current LLMs perform in the domain of climate change. Our evaluation of some widely-used open-source LLMs revealed systematic patterns: few-shot prompting generally improves performance, but certain text classification tasks such as claim detection or climate-specific NER remain challenging. We hope that ClimateEval will serve as a valuable resource for the NLP community and facilitate future research on evaluating and improving LLMs for climate change-related applications.

## Limitations

We acknowledge that ClimateEval is currently limited to English, which restricts its applicability to multilingual climate discourse. Due to computational constraints our evaluation focuses on mid-sized open-source models, ranging from 2B

Task	7B		13B	
	0	5	0	5
CDP-QA-Corp.	<b>0.26</b>	<b>0.02</b>	<b>0.20</b>	<b>0.04</b>
CDP-QA-States	<b>0.23</b>	<b>0.02</b>	<b>0.13</b>	<b>0.04</b>
CDP-QA-Cities	<b>0.22</b>	<b>0.03</b>	<b>0.13</b>	<b>0.05</b>
CDP-Topic-Cities	-0.05	0.00	-0.03	0.00
ClimaText Sent. Clf.	-0.00	-0.06	-0.10	-0.02
Climate NER	-0.03	-0.04	-0.04	0.00
Climate Commit.	<b>0.05</b>	-0.03	-0.07	<b>0.04</b>
Climate Detection	<b>0.01</b>	<b>0.07</b>	-0.06	<b>0.08</b>
Climate Eng	<b>0.16</b>	0.00	-0.00	0.00
Climate-FEVER	-0.15	<b>0.11</b>	<b>0.07</b>	<b>0.14</b>
Climate Sentiment	<b>0.17</b>	-0.06	-0.03	-0.06
Climate Specificity	-0.06	<b>0.01</b>	<b>0.02</b>	<b>0.04</b>
Climate Stance	-0.09	-0.10	-0.01	-0.06
Env. Claims	<b>0.19</b>	-0.05	-0.02	<b>0.03</b>
Exeter Claim	<b>0.09</b>	<b>0.07</b>	<b>0.14</b>	0.03
Exeter Sub-claim	<b>0.19</b>	<b>0.12</b>	<b>0.19</b>	<b>0.16</b>
Guardian Body	<b>0.02</b>	<b>0.06</b>	<b>0.02</b>	<b>0.24</b>
Guardian Title	<b>0.06</b>	<b>0.07</b>	<b>0.07</b>	<b>0.08</b>
Net-Zero Reduction	-0.24	-0.01	<b>0.04</b>	-0.01
PIRA w/ Ctx.	<b>0.12</b>	<b>0.17</b>	<b>0.06</b>	<b>0.06</b>
PIRA w/o Ctx.	<b>0.23</b>	<b>0.23</b>	<b>0.12</b>	<b>0.15</b>
SciDCC Title	-0.02	-0.07	<b>0.08</b>	<b>0.04</b>
SciDCC Title Sum.	<b>0.04</b>	-0.10	<b>0.08</b>	0.00
SciDCC Title Body	-0.00	-0.02	<b>0.08</b>	<b>0.01</b>
TCFD Recommend.	-0.08	-0.02	<b>0.05</b>	<b>0.07</b>
Average	<b>0.05</b>	<b>0.01</b>	-0.04	<b>0.04</b>

Table 4: Comparison of ClimateGPT against its’ base Llama models in **0**-shot and **5**-shot settings. The values represent the difference where positive values (highlighted in bold) indicate better performance of ClimateGPT over Llama.

to 70B parameters, with larger models tested under 4-bit quantization. This may introduce some performance degradation and may not reflect their optimal performance. Additionally, commercial models such as GPT-4 and Claude are not included in our experiments due to budget constraints. Finally, the benchmark largely consists of  $n$ -way classification tasks, with the only exception of Climate-Change NER. This was partly driven by available climate-relevant datasets, which are predominantly classification-oriented. Future work should focus on enabling assessment of generative tasks such as information extraction, text generation or multi-modal classification.

## Acknowledgments

Swedish Research Council Vetenskapsrådet grant nos. 2022-03448, 2022-06599 and 2022-02909. European Commission’s Horizon Europe research and innovation programme European Research Council grant no. 101112727.

## References

- Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Michael M. Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain V. Costes, Sergi Blanco-Cuaresma, Kelly Lockhart, Thomas Allen, Felix Grezes, Megan Ansdell, Alberto Accomazzi, Yousef El-Kurdi, Davis Wertheimer, Birgit Pfitzmann, Cesar Berrospi Ramis, Michele Dolfi, Rafael Teixeira De Lima, Panagiotis Vagenas, S. Karthik Mukkavilli, Peter W. J. Staar, Sanaz Vahidinia, Ryan McGranaghan, and Tsengdar J. Lee. 2024. **INDUS: Effective and Efficient Language Models for Scientific Applications**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 98–112, Miami, Florida, US. Association for Computational Linguistics.
- Julia Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2023. How Cheap Talk in Climate Disclosures Relates to Climate Initiatives, Corporate Emissions, and Reputation Risk. Working paper, Available at SSRN 3998435.
- 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):1–12.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Builan, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. *arXiv preprint arXiv:2012.00614*.
- May El Barachi, Manar AlKhatib, Sujith Mathew, and Farhad Oroumchian. 2021. **A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change**. *Journal of Cleaner Production*, 312:127820.
- Michael Fore, Simranjit Singh, Chaehong Lee, Amritanshu Pandey, Antonios Anastasopoulos, and Dimitrios Stamoulis. 2024. **Unlearning Climate Misinformation in Large Language Models**. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 178–192, Bangkok, Thailand. Association for Computational Linguistics.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. **A framework for few-shot language model evaluation**.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Amirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Alionsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junting Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharani Narang, Sharath Raparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal

Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonquet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Chingsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khan-delwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madiam Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov,

Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghatham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaoqian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaodu Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The Llama 3 Herd of Models](#). Preprint, arXiv:2407.21783.

Rodney Michael Kinney, Chloe Anastasiades, Russell Author, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarion, Madeleine van Zuylen, and Daniel S. Weld. 2023.

- The Semantic Scholar Open Data Platform. *ArXiv*, abs/2301.10140.
- Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *Preprint*, arXiv:1910.09700.
- Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A Survey on Deep Learning for Named Entity Recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.
- Ni Li, Shorouq Zahra, Mariana Brito, Clare Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, et al. 2024. Using LLMs to Build a Database of Climate Extreme Impacts. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 93–110.
- Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Carl Yang, and Liang Zhao. 2023. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey.
- Mariana Madruga de Brito, Jan Sodoge, Heidi Kreibich, and Christian Kuhlicke. 2025. Comprehensive assessment of flood socioeconomic impacts through text-mining. *Water Resources Research*, 61(1):e2024WR037813.
- Prakamya Mishra and Rohan Mittal. 2021. Neural-NERE: Neural Named Entity Relationship Extraction for End-to-End Climate Change Knowledge Graph Construction. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.
- Paulo Pirozelli, Marcos M José, Igor Silveira, Flávio Nakasato, Sarajane M Peres, Anarosa AF Brandão, Anna HR Costa, and Fabio G Cozman. 2024. Benchmarks for pirá 2.0, a Reading Comprehension Dataset about the Ocean, the Brazilian Coast, and Climate Change. *Data Intelligence*, 6(1):29–63.
- Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards Answering Climate Questionnaires from Unstructured Climate Reports. *Preprint*, arXiv:2301.04253.
- Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental Claim Detection. *arXiv preprint arXiv:2209.00507v4*, arXiv:2209.00507.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussonot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Letícia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving

Open Language Models at a Practical Size. *Preprint*, arXiv:2408.00118.

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. *Preprint*, arXiv:2401.09646.

Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A Multi-task Model for Sentiment Aided Stance Detection of Climate Change Tweets. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 854–865.

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.

Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2021. ClimaText: A Dataset for Climate Change Topic Detection.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jian-hong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tian-hao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Francisco Zanartu, Yulia Otmakhova, John Cook, and Lea Frermann. 2024. Generative Debunking of Climate Misinformation. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 46–62, Bangkok, Thailand. Association for Computational Linguistics.

Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Devasier, and Chengkai Li. 2024. Granular Analysis of Social Media Users’ Truthfulness Stances Toward Climate Change Factual Claims. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 233–240.

## A Models

Table 5 lists the baseline models used in our paper, along with their corresponding repository names on <https://huggingface.co/>.

Model Name	HuggingFace Repository
Gemma-2-2B	google/gemma-2-2b-it
Qwen-2.5-7B	Qwen/Qwen2.5-7B
Llama-2-7B	meta-llama/Llama-2-7b-chat-hf
Llama-2-13B	meta-llama/Llama-2-13b-chat-hf
ClimateGPT-7B	climategpt/climategpt-7b
ClimateGPT-13B	climategpt/climategpt-13b
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
Mistral-24B	mistralai/Mistral-24B-Instruct
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct

Table 5: HuggingFace repository names of the baseline models used in our evaluation.

## B CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO<sub>2</sub>eq/kWh. A cumulative of 240 hours of computation was performed on hardware of type A100 SXM4 80 GB (TDP of 400W). Total emissions are estimated to be 41.47 kgCO<sub>2</sub>eq of which 0 percent were directly offset. Estimations were conducted using the **MachineLearning Impact calculator** presented in Lacoste et al. (2019).

## C Label distribution

To provide further insights into the datasets, we visualize the label distribution for each test set in Figure 3. Given the wide size range of the test sets (between 300 and 55872), we present the normalized label distributions, where each stack in a bar represents the percentage of a label within the corresponding task’s test set.

Dataset (Source)	Task(s)	Shortened identifier
ClimaText (Varini et al., 2021)	Sentence classification	ClimaText Sent. Clf.
Climate-Stance (Vaid et al., 2022)	Stance classification	Climate-Stance
Climate-Eng (Vaid et al., 2022)	Topic classification	Climate-Eng
Climate-FEVER (Diggelmann et al., 2020)	Claim verification	Climate-FEVER
SciDCC (Mishra and Mittal, 2021)	Topic classification by Title Topic classification by Title & Summary Topic classification by Title & Body	SciDCC Title SciDCC Title Sum. SciDCC Title Body
CLIMA-CDP (Spokoyny et al., 2023)	QA-Cities (answer relevance) QA-Corporations (answer relevance) QA-States (answer relevance) QA-Topic-Cities (topic classification)	QA-Cities QA-Corps. QA-States QA-Topic-Cities
PIRA 2.0 MCQ (Pirozelli et al., 2024)	PIRA with Context PIRA without Context	PIRA w/ Ctx. PIRA w/o Ctx.
Exeter Misinformation (Coan et al., 2021)	Claim Detection Sub-claim Detection	Exeter Claim Exeter Sub-claim
Climate-Change NER (Bhattacharjee et al., 2024)	Entity recognition	Climate NER
CheapTalk (Bingler et al., 2023)	Climate Detection Climate Sentiment Climate Commitment Climate Specificity TCFD Recommendations	Climate Commit. TCFD Recommend.
Net-Zero Reduction (Schimanski et al., 2023)	Paragraph classification	Net-Zero Reduction
Environmental Claims (Stammabach et al., 2023)	Sentence classification	Env. Claims
Guardian Climate News Corpus	Topic classification by Title Topic classification by Body	Guardian Title Guardian Body

Table 6: Shortened task identifiers (if they exist) for each task presented in the ClimateEval benchmark. These shortened task names are used in tables presenting the results for the purpose of saving space.

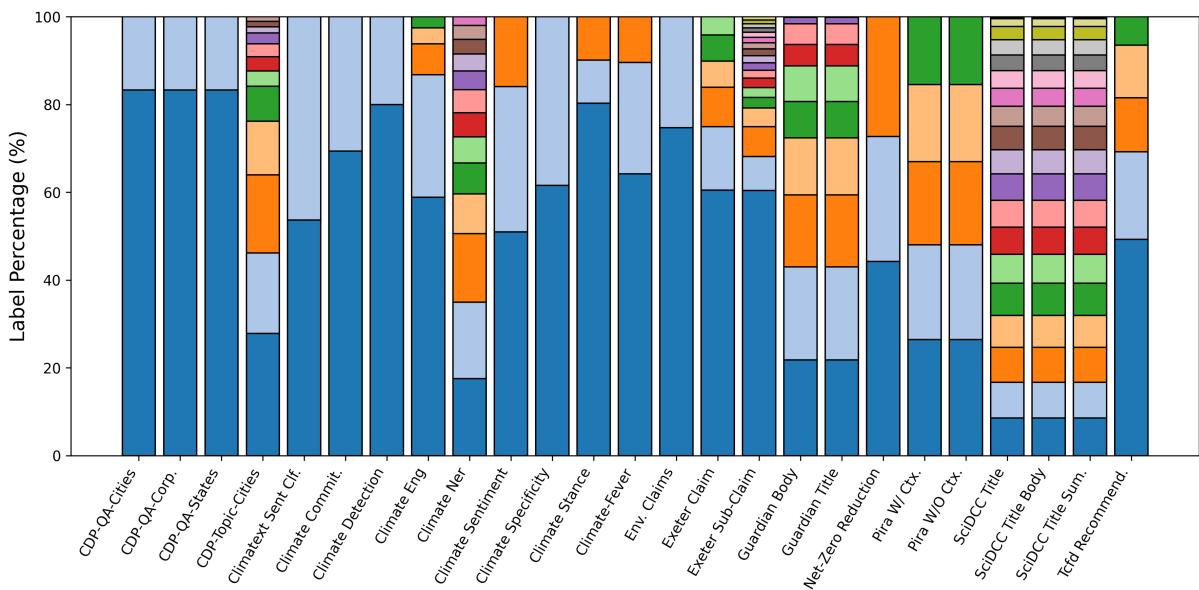


Figure 3: The normalized distribution of labels in each task across all sets (train, development, and test).