ClimaBench: A Benchmark Dataset For Climate Change Text Understanding in English

Tanmay Laud*
UCSD*
tlaud@ucsd.edu

Daniel Spokoyny*
CMU
dspokoyn@andrew.cmu.edu

Tom Corringham
SIO
tcorringham@ucsd.edu

Taylor Berg-Kirkpatrick
UCSD
tberg@eng.ucsd.edu

Abstract

The topic of Climate Change (CC) has received limited attention in NLP despite its real world urgency. Activists and policy-makers need NLP tools in order to effectively process the vast and rapidly growing textual data produced on CC. Their utility, however, primarily depends on whether the current state-of-the-art models can generalize across various tasks in the CC domain. In order to address this gap, we introduce Climate Change Benchmark (ClimaBench), a benchmark collection of existing disparate datasets for evaluating model performance across a diverse set of CC NLU tasks systematically. Further, we enhance the benchmark by releasing two large-scale labelled text classification and question-answering datasets curated from publicly available environmental disclosures. Lastly, we provide an analysis of several generic and CC-oriented models answering whether fine-tuning on domain text offers any improvements across these tasks. We hope this work provides a standard assessment tool for research on CC text data.

1 Introduction

There is an ever growing body of text-based climate documents that contain vital information on carbon emissions, impacts, and risks – for example, firms announce their emissions reduction targets or cities disclose their water risks and potential exposure to drought. These documents include corporate ESG (Environmental, Social, and Governance), climate assessments, climate legislation, regulatory filings, and are produced by corporations, cities, states, and national governments. Although, there have been massive efforts by international organizations to standardize climate reporting and disclosures, the vast majority of data is still non machine-readable. Natural language processing has potential to help researchers and policy-makers search, extract and gain insights from climate texts. Several works have introduced small-scale climate-specific datasets for detecting relevance to climate (Leippold and Varini, 2020), identifying stance detection (Vaid et al., 2022b) and fact-checking (Leippold and Diggelmann, 2020) of social media claims. Other works have studied the effect of further pretraining of transformers using in-domain data such as news articles, academic papers, and climate reports (Luccioni et al., 2020; Webersinke et al., 2021).

However, the use of proprietary datasets for both pretraining as well as evaluation makes it difficult to effectively track progress in this field. To help unify the disparate efforts of applying NLP technologies to climate-related problems we propose CLIMABENCH, a benchmark for training and evaluating NLP models on classification and question-answering tasks pertaining to climate texts. Benchmarks have been an effective tool for the general NLP community (Wang et al., 2018; Srivastava et al., 2022; Liang et al., 2020) as well as for specific areas such as legal (Chalkidis et al., 2022) or biomedical (Lee et al., 2020) domains.

We introduce two new large scale datasets, CLIMA-CDP and CLIMA-INSURANCE, which are based on publicly accessible climate questionnaires of cities, states and corporations. In contrast with existing datasets, these questionnaires have an order of magnitude more examples and the examples themselves are longer. Further, CLIMA-QA uses a highly structured questionnaire: with hundreds of unique question types across a multitude of topics such as transportation, water security, and governance. Additionally, we collate five existing climate datasets and, to the best of our knowledge, conduct the first evaluation on the SCIDCC (Mishra and Mittal, 2021) topic classification dataset.

A common technique to adapt large language

* Equal Contribution

Dataset	Task Type	Domain	# Train	# Dev	# Test	# Classes
CLIMA-INSURANCE (Ours)	Binary Classification	NAIC	13.8K	1.6K	1.7K	2
CLIMA-INSURANCE+ (Ours)	Multi-class Classification	NAIC	13.7K	1.7K	1.7K	8
CLIMA-CDP (Ours)	Topic Classification	CDP Cities	46.8K	8.7K	8.9K	12
CLIMA-QA (Ours)	QA Ranking	CDP Cities	48.2K	8.5K	9.3K	294
		CDP States	8.7K	0.9K	1.1K	132
		CDP Corporations	34.5K	3.6K	4.9K	43
CLIMATEXT (Leippold and Varini, 2020)	Binary Classification	Wikipedia, 10-K	6K	0.3K	1.6K	2
CLIMATESTANCE (Vaid et al., 2022a)	Ternary Classification	Twitter	3K	0.3K	0.3K	3
CLIMATEENG (Vaid et al., 2022a)	Multi-class Classification	Twitter	3K	0.3K	0.3K	5
CLIMATEFEVER (Leippold and Diggelmann, 2020)	Fact-Checking	Wikipedia	-	-	1.5K	3
SCIDCC (Mishra and Mittal, 2021)	Topic Classification	Science Daily	9.2K	1.1K	1.1K	20

Table 1: General Statistics of the CLIMABENCH datasets

	Text	Class / Question
CLIMA-INSURANCE+	...Each year Aflac reports its US operations Scope 1 and Scope 2 emissions to the Carbon Disclosure Project. Since 2007, Aflac's owned facilities in terms of square feet have increased by more than 10% while total Scope 1 and 2 CO2e emissions have significantly decreased compared to 2007 emissions...	Question Type 1 / Yes
CLIMA-CDP	...These Plans must include management of CD&E waste, both through on-site recycling and re-use and on-site waste processing prior to disposal.Westminster will contribute to the London Plan target of net self-sufficiency (managing 100% of London's waste within London) by 2026 by planning for Westminster's apportionment targets...	Governance and Data Management
CLIMA-QA	Flooding from sea level rise will damage building and roads in the coastal neighborhoods of the city. Flooding also represents a risk to major transportation hubs infrastructure in the region. Coastal flooding can have a long-term effect on major industrial and commercial activities along the coastal areas of the city as well as damage urban forestry and local natural biodiversity.	Please describe the impacts experienced so far, and how you expect the hazard to impact in the future.
SCIDCC	...In the past, wave breaking has been tracked by so-called "whitecap coverage," in which still or video imagery was used to statically identify ocean whitecaps and the corresponding surface coverage by breaking waves. But those measurements suffered...	Hurricanes and Cyclones

Table 2: Examples (pairs of inputs and outputs) for some of the newly introduced datasets in CLIMABENCH

models (LLM) to new tasks has been to pretrain on in-domain texts. In the CC domain, datasets of corporate sustainability reports, climate news articles, climate-related research abstracts and financial reports from EDGAR[1] have been used for pretraining (Luccioni et al., 2020; Webersinke et al., 2021). However, certain alternate sources of climate text data are proprietary and models trained on this data are generally unavailable for research. For example, the Global Reporting Initiative (GRI), a non-governmental organization, once hosted over 63,000 corporate sustainability reports which are now inaccessible[2].

Our experiments on CLIMABENCH suggest that state-of-the-art LLMs provide satisfactory performance when finetuned on the downstream CC tasks, but there is still substantial room for improvement. Further, there was no single model that dominated the rest on all of the classification tasks. Beyond evaluation of existing models we also show that training models on our new benchmark can be directly applied to a real-world climate focused downstream task. Specifically, using the best performing model trained on our new benchmark, we predict relevant questions for text segments from U.S. State climate action plans which is a completely different type of CC document. Out of the 500 predictions judged by a climate change researcher 71% were relevant and 35.6% highly relevant.

We hope that CLIMABENCH lowers the barrier for the NLP community to conduct research on cli-

[1]U.S. Securities and Exchange Commission Electronic Data Gathering, Analysis, and Retrieval (EDGAR) Database of public company filings.

[2]https://www.globalreporting.org/how-to-use-the-gri-standards/register-your-report/

mate related text tasks. We release our benchmark[3] and open-source our trained models[4] to encourage researchers to extend our existing datasets and contribute new ones.

2 CLIMABENCH Tasks and Datasets

2.1 New Datasets

CDP (formerly the Carbon Disclosure Project[5]) is an international non-profit organisation that helps companies and cities disclose climate risks, low carbon opportunities and environmental impact. In 2021, over 14,000 organizations disclosed their environmental information via CDP. The annual National Association Of Insurance Commissioners (NAIC)[6] Climate Risk Disclosure Survey[7] is a U.S. insurance regulation tool where insurers file non-confidential disclosures of their assessments and management of climate-related risks. The purpose of the survey is to enhance transparency about how insurers manage climate-related risks and opportunities and enable better-informed collaboration on climate-related issues. It consists of eight questions and the expected response is a Yes/No along with an explanation (see Table 2 for example).

We release two new large-scale datasets with four tasks, based on publicly available surveys released by CDP and NAIC. **CLIMA-INSURANCE** and **CLIMA-CDP** are for text classification tasks. **CLIMA-QA** is a question answering task based on the CLIMA-CDP dataset. Table 2 lists a few interesting examples from the newly introduced datasets.

Practical Implications. Recently (2022), NAIC eliminated mandatory Insurance questionnaires to lessen the burden of compliance, allowing for free text report submissions. However, as the structured questionnaires provided an organized assessment of sector-wise exposure to climate risks, NAIC now invests substantial manual effort in mapping free text response segments to the original questions. A model trained on our new CLIMA-INSURANCE dataset could drastically help in this effort as this task very closely matches the real-world problem faced by NAIC. NAIC is just one organization out of many that touch on various aspects related to societal robustness to climate change. If we improve efficiency in processing similar reports in other domains affecting climate change (e.g. economic and ecological reporting), we can better inform policy decisions.

Similarly, Climate Watch, a platform managed by the World Resources Institute, lets users analyze and compare the Nationally Determined Contributions (NDCs) under the Paris Agreement and discover how countries can leverage their climate goals to achieve their sustainable development objectives. The data curation involves manually mapping the NDCs to their corresponding Sustainable Development Goals, a task which has a strong overlap with the categories in our CLIMA-CDP task.

2.2 Data Processing

The overall statistics of each dataset are given in Table 1, the token length distribution is given in Appendix Table 9 and details are explained below.

CLIMA-INSURANCE We create two tasks based on the responses to the NAIC survey, first being a binary Yes/No classification task (CLIMA-INSURANCE) and the other one being a question type classification problem (CLIMA-INSURANCE+). The data is web scraped from all the annual survey responses between 2012-2021 giving a total of 17K labelled passages. We remove the first sentence in each response as it contains obvious markers (like "Yes, we do X." or "No, we do not participate in Y.") and create splits for training, validation and testing. The splits of the two datasets are different since we stratify by the class label in order to fairly balance the classes across the splits.

CLIMA-CDP The CDP survey responses fall into 3 buckets, namely Cities, States and Corporations, each with a different questionnaire with varying number of questions. We filter out non-English, short (less than 10 words) and duplicate responses. The Cities dataset is the largest by volume and consists of parent sections and subsections. We transform these sections into 12 broad categories with the help of our climate change researcher to curate the labelled data called CLIMA-CDP for the years 2018-2021 (Appendix Table 8). The goal is to classify CC data into relevant categories. The 12-category mapping creates a more parsimonious set of labels which can be compared to other studies and reduces noise in classification (The original CDP section labels have changed slightly in the

[3] https://github.com/climabench/climabench

[4] https://huggingface.co/climabench/miniLM-cdp-all

[5] https://www.cdp.net/

[6] http://www.insurance.ca.gov/0250-insurers/0300-insurers/0100-applications/ClimateSurvey/

[7] https://interactive.web.insurance.ca.gov

period 2018-2021). The train, development and test splits are stratified by the organization so that some organization responses are not seen during training.

CLIMA-QA Although the categories in CLIMA-CDP are valuable for CC topic classification, they restrict the downstream application to finite classes. Hence, we extend our focus on the actual question response pairs within the surveys to set up a more nuanced and challenging QA dataset called CLIMA-QA. Since there is no explicit output class, the hope is that models finetuned on this task should be able to generalise to unseen question-answer pairs, and we conduct extensive experiments to study this effect. The pre-processing steps are the same as that of CLIMA-CDP. We curate three different subsets: **CDP-CITIES, CDP-STATES and CDP-CORPORATION** (Table 1) where the splits are stratified by the organizations. They have 294, 132 and 43 unique questions respectively.

2.3 Collating existing datasets

Further, we collate existing CC related text datasets, described in detail below.

CLIMATEXT is a dataset for sentence-based climate change topic detection (Leippold and Varini, 2020). Each sentence is labelled indicating whether it talks about climate change or not. Sentences were collected from different sources: Wikipedia, the U.S. Securities and Exchange Commission (SEC) 10Kfiles, which are annual regulatory filings in which listed companies in the US are required to self-identify climate-related risks that are material to their business, and a selection of climate-change claims collected from the web.

CLIMATESTANCE and **CLIMATEENG** Vaid et al. (2022a) extracted Twitter data consisting of 3777 tweets posted during the 2019 United Nations Framework Convention on Climate Change. Each tweet was labelled for two tasks, stance detection and categorical classification. For the stance towards climate change prevention, the authors labelled each tweet as In Favour, Against or Ambiguous. For categorical classification, the five classes are Disaster, Ocean/Water, Agriculture/Forestry, Politics, and General.

CLIMATEFEVER (Leippold and Diggelmann, 2020) adopts the methodology of FEVER, the largest dataset of artificially designed claims, for real-life claims on climate change collected online. It consists of 1,535 claims and each claim is mapped with five relevant evidence passages from Wikipedia giving a total of 7675 claim-evidence pairs. The labels for the evidences are SUPPORTS, REFUTES or NOT_ENOUGH_INFO. The dataset is also part of the recent BEIR benchmark for information retrieval (Thakur et al., 2021).

SCIDCC (Mishra and Mittal, 2021) The Science Daily Climate Change (SCIDCC) dataset is curated by web scraping news articles from the Science Daily (SD) website. It contains around 11k news articles from 20 categories relevant to climate change, where each article comprises of a title, summary, and a body. Some of the major categories are Earthquakes, Pollution, Hurricanes and Cyclones. We propose to use this dataset for the first time as a category classification task for two reasons. Firstly, the SD news articles are relatively more scientific as compared to other online news. Secondly, the average document length is around 500-600 words with a maximum of roughly 2.5k words, which is significantly longer than other existing public CC datasets.

3 Models

We experiment with simple baselines (Majority class and Random class), Linear models and pretrained Transformer-based classfiers. The Random classifier uniformly samples a label for each input while the Majority one always outputs the most frequent label.

3.1 Linear Models

We assess Support Vector Machines (SVM) with linear kernel as one of the baseline models. The input features are the TF-IDF vectors with word n-grams in the range (1,3). The linear models are only trained on the text classification tasks.

3.2 Pretrained Transformers

We examine Transformer-based (Vaswani et al., 2017) pre-trained language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). These models are first pretrained on very large unlabeled corpora on self-supervised learning tasks like masked language modeling or next sentence prediction. Then, they are fine-tuned on smaller task-specific labelled datasets, after adding task-oriented layers. We also examine distilled versions like DistilRoBERTa (Sanh et al., 2019), longer context models like Longformer (Beltagy et al., 2020), and domain specific models like ClimateBERT (We-

bersinke et al., 2021) and SciBERT (Beltagy et al., 2019). This helps us contrast the effects of model architecture, input length and in-domain pretraining on downstream tasks. Thus, we fine-tune and evaluate the performance of the publicly available models. We provide more details about models in Appendix Section A.3 and Table 10.

MiniLM For QA (Wang et al., 2020) is an uncased distilled model released by Microsoft based on an in-house pre-trained UniLM v2 model in BERT-Base size. Reimers and Gurevych (2019) separately finetuned this model on the MSMARCO (Campos et al., 2016) QA ranking dataset and achieved state of the art performance with roughly 2.0x speedups over traditional BERT based rankers. We therefore consider this QA finetuned model (MSMARCO-MiniLM) as a strong baseline for our fine-tuning and transfer learning experiments on CLIMA-QA.

3.3 Task specific finetuning

Text Classification For the text classification datasets like CLIMATEXT, CLIMATESTANCE, CLIMATEENG, SCIDCC and CLIMA-CDP, we pass each document through the model and then apply linear head on top of the $[CLS]$ token representation of the model which is the defacto standard for finetuning transformer based language model classifiers. We apply a softmax layer on top of the linear head to get the probability distribution over classes. For SCIDCC, we concatenate the text fields and provide a train, validation and test split (80%, 10%, 10%) stratified by the categories.

Fact Checking For CLIMATEFEVER, we convert each data point (the claim and 5 evidences with labels) into five input pairs. Each pair consists of the claim and one of the five candidate evidences, separated by the special $[SEP]$ delimiter token. The output representation of the $[CLS]$ token of each pair is passed to a linear layer followed by a softmax layer to obtain the output probabilities. We randomly sample train, dev and test samples in 8:1:1 ratio, stratified by the labels.

Question Answering For CLIMA-QA, we randomly sample 5 negative QA pairs for each positive QA pair in the dataset (1:5 ratio). Similar to the fact-checking setup, we pass each pair through the pretrained model, separated by the $[SEP]$ token. This formulation more closely aligns with question answer ranking setup trained on datasets like MSMARCO (Campos et al., 2016). We specifically consider the early-interaction encoder (or cross-encoder) setup to simultaneously model the context of the question and answer, allowing for the self-attention to interact between the question and answer throughout the entire model. We narrow down to two models, the MSMARCO finetuned MiniLM, and the ClimateBERT model to study the effects of fine-tuning and transfer learning on the three subdomains: CDP-CITIES, CDP-STATES and CDP-CORPORATION.

4 Experiments

4.1 Setup

We use the Scikit-learn API (Pedregosa et al., 2011) for the simple classifiers (Random and Majority class) and TFIDF-based linear SVM models. We grid-search the hyper parameters for SVM with 5-fold validation (Table 12). For all the pre-trained models, we use publicly available Hugging Face (Wolf et al., 2020) checkpoints.[8] We use Adam optimizer (Kingma and Ba, 2015) with learning rate of 5e-5 (linear warmup ratio of 0.1, weight decay of 0.01) for 10 epochs with early stopping based on performance on development data (macro-F1 score). We use mixed precision (fp16), gradient checkpointing and gradient accumulation steps of 2 to train models efficiently on the limited compute (Appendix A.1). For the Longformer, we use the default settings (windows of 512 tokens and a single global $[CLS]$ token). The training batch size is set to 32. We use weighted cross-entropy loss for all the pretrained transformer models with class balanced weights. We truncate the input text when it exceeds the maximum input length of the model and pad otherwise.

4.2 Evaluation

The classification models are evaluated using the macro-averaged F1 score on the development and test sets. We use macro-average since the datasets are imbalanced and all classes are equally important. We do not evaluate linear models on fact-checking or QA as the heterogeneity of the input in these tasks do not align with the linear setup.

For the ranking task, we consider the Mean Reciprocal Rank at k (MRR@k) scores for the top k

[8] We use the *-base configuration of each pre-trained model, i.e., 12 Transformer blocks, 768 hidden units, and 12 attention heads. For ClimateBERT we report scores for the F variant model on Huggingface. For the QA Cross-encoder, we use the MiniLM (12 layer, 384 hidden-unit) finetuned on MSMARCO available at https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

Models	CLIMA-INSURANCE	CLIMA-INSURANCE+	CLIMA-CDP	CLIMA-TEXT	CLIMATE-STANCE	CLIMATE-ENG	SCIDCC	CLIMATE-FEVER	Average
Majority Classifier	0.4058	0.0411	0.0365	0.4208	0.2968	0.1383	0.0079	0.2608	0.2010
Random Classifier	0.4940	0.1214	0.0645	0.4686	0.2552	0.1671	0.0505	0.3062	0.2409
TFIDF + Linear SVM	0.8150	**0.8600**	0.5834	0.8339	0.4292	0.5181	0.4802	-	-
BERT	0.8396†	0.8457	0.6464†	0.8704†	0.5537†	0.7178	0.5474†	0.6247†	0.7057†
RoBERTa	**0.8445**	0.8561†	**0.6522**	0.8597	**0.5969**	**0.7458**	0.5290	0.6074	**0.7114**
DistilRoBERTa	0.8263	0.8438	0.6361	0.8606	0.5251	0.7233†	0.5113	0.6154	0.6927
Longformer	0.8301	0.8435	0.6403	**0.8780**	0.3468	0.7228	**0.5479**	0.6082	0.6772
SciBERT	0.8260	0.8443	0.6362	0.8329	0.4867	0.7050	0.5183	**0.6268**	0.6845
ClimateBERT	0.8214	0.8480	0.6424	0.8514	0.5284	0.7183	0.5297	0.6154	0.6944

Table 3: Macro F1 Scores on the Classification Datasets. **Bold** and † indicate first and second highest performing model respectively. RoBERTa scores the best on average followed by BERT and ClimateBERT.

items returned by a model. MRR, a popular metric used in the Information Retrieval field, is the average of the reciprocal ranks of results for a sample of queries where the relevance grading is binary (Yes/No). One notable distinction is that we retrieve the top questions for a given document as opposed to traditional search where the top answers are returned. We explain the rationale behind this peculiar design choice in Section 6 where we conduct a pilot study with our climate expert. First, we construct the baseline with BM25 (Robertson and Zaragoza, 2009) and MSMARCO-MiniLM, where we evaluate on the three test sets CDP-CITIES, CDP-STATES, CDP-CORPORATION without any finetuning. Next, we finetune on the largest of the three corpora, CDP-CITIES, and then evaluate on all three datasets to analyse the transfer learning capability on States and Corporations. Similarly, we train and test on States and Corporations. Lastly, we train and test the best model on all three datasets combined. We also fine-tune ClimateBERT as a second baseline.

4.2.1 Efficiency

Further, we report the compute efficiency statistics like average training time and average training samples per second for each model in Appendix Table 11. These numbers are hardware dependent but provide a relative efficiency comparison between the models. The details about the compute and approximate CO_2 Emission usage is provided in A.2.

5 Results

5.1 Classification and Fact-Checking

Table 3 reports the F1-macro on the test sets for the classification and fact-checking tasks in CLIMABENCH. We also give the average of the scores on all tasks for each model. There is no single model that does the best across the board, but RoBERTa is a clear winner as it beats the other baselines on four out of eight tasks. BERT falls second in line when we look at average scores. Longformer scores the highest on SCIDCC, perhaps due to the model being able to attend to more number of tokens available in the longer documents in this task. To our surprise, it also does well on CLIMATEXT which has comparatively shorter text. One reason for this could be that the sparse attention within Longformer may be having a regularizing effect. SciBERT ranks on the top for CLIMATEFEVER and this could be because the evidences in the task have structure and style similar to the text in scientific papers used for pretraining the model. ClimateBERT and the model it was warm-started from, DistilRoBERTa, are very similar in performance. DistilRoBERTa beats ClimateBERT on CLIMA-INSURANCE, CLIMATEXT, and CLIMATEENG. Although this raises concerns on the quality of pretraining data used for ClimateBERT, we cannot say for sure, since neither the data nor the code is publicly available for diagnosis. Overall, the transformer models have significantly better gains over linear ones except on CLIMA-

	CDP-CITIES				CDP-STATES			CDP-CORPORATION	
Model	**MRR@10**	**MRR@100**	**MRR@200**	**MRR@All**	**MRR@10**	**MRR@100**	**MRR@All**	**MRR@10**	**MRR@All**
No Finetuning on CDP									
BM25	0.055	0.075	0.076	0.077	0.084	0.104	0.105	0.153	0.180
MiniLM	0.099	0.116	0.117	0.118	0.120	0.141	0.142	0.320	0.342
Finetuned on CDP									
	In-Domain				In-Domain			In-Domain	
ClimateBERT	0.331	0.344	0.344	0.344	0.422	0.431	0.431	0.753	0.754
MiniLM	**0.366**	**0.378**	**0.378**	**0.378**	0.482	0.491	0.491	**0.755**	**0.757**
Best Model Finetuned on all									
MiniLM	0.352	0.364	0.364	0.364	**0.489**	**0.497**	**0.497**	0.745	0.747

Table 4: MRR@k scores for BM25, ClimateBERT and MSMARCO-MiniLM on the three subsets of CLIMA-QA. Models finetuned and evaluated on same subset fall under In-Domain.

	CDP-STATES			CDP-CORPORATION	
Model	**MRR@10**	**MRR@100**	**MRR@All**	**MRR@10**	**MRR@All**
No Finetuning					
BM25	0.084	0.104	0.105	0.153	0.180
MiniLM	0.120	0.141	0.142	0.320	0.342
Finetuned on CDP-CITIES					
	Transfer			Transfer	
ClimateBERT	0.298	0.314	0.314	0.465	0.477
MiniLM	**0.353**	**0.366**	**0.366**	**0.489**	**0.500**

Table 5: MRR@k scores for BM25, ClimateBERT and MSMARCO-MiniLM on the Transfer experiments. Models are finetuned on CDP-CITIES and evaluated on States and Corporations.

INSURANCE+ where the TFIDF+SVM model is superior. It shows that simple word co-occurence statistics are enough for certain tasks and deep language models might not be the right solution in such cases.

5.2 Question Answer Ranking

We analyze and compare the results of our in-domain and transfer experiments on CLIMA-QA. Table 4 and Table 5 summarize the MRR@k results for the in-domain and transfer experiments respectively. MSMARCO-MiniLM beats ClimateBERT on all the subsets, for both in-domain and transfer learning. Both models do significantly better when fine-tuned on CDP-CITIES and evaluated on States and Corporations. This indicates that models trained on this dataset possess good transfer learning capabilities. The in-domain training boosts performance further as we see for States and Corporations. Lastly, the best performing model, MiniLM, when finetuned on all three subsets, achieves comparable performance on Cities and Corporations while ranking highest on States.

6 Extrinsic Evaluation

Beyond CDP. States and Companies have been utilizing sustainability reporting frameworks to help guide their non-financial reporting efforts. These frameworks are useful given the lack of uniform regulations on nonfinancial reporting. However, the variety of such frameworks can be daunting. For example, CDP aims to capture companies' environmental performance data, whereas the Task Force on Climate-related Financial Disclosures[9](TCFD) guidelines are intended to encourage companies to align their climate-related risk disclosures with investors' needs. Automatic alignment of text to relevant sections of the disclosures reduces the effort significantly. To motivate this usecase of the CLIMA-QA task, we conduct an experiment that involves populating the CDP survey using Climate Action Plans (CAPS).

Utilizing CAPs. Disclosure of carbon accounting, climate risks and green targets is vital information for researchers and policy-makers to craft well-informed climate policy rules and regulations. In addition to these reports, cities, states, and corporations may publish their CAP or sustainability reports. The CAPS include quantitative data, such as emission values or renewable electricity generation capacity, and qualitative data such as specific policy interventions across different sectors. Populating CDP surveys from new CAPs allows for consistent comparisons of CAPs to existing datasets which could be used to compare strategies, identify gaps, or rank jurisdictions on the content and level of ambition in their stated plans. We consider the scenario where the State's CAP is available but their CDP report is not. The researcher's aim is to

[9]https://www.fsb-tcfd.org

	prec@1	prec@2	prec@3	prec@4	prec@5
Relevant	63.0	67.0	68.6	69.5	71.0
Highly Relevant	30.0	32.0	32.3	32.5	35.6

Table 6: Precision@K: We report the fraction of items in the top K ranked retrievals that are either marked as highly relevant, or at least relevant, averaged across text examples. Relevance judgements for were performed manually by an expert annotator.

pull together the relevant portions of the CAP to construct the missing CDP. Typically the CAPs are much longer (∼100 pages) and more comprehensive than any particular disclosure report. Our aim is to measure the effectiveness of our model to assist in this task.

6.1 Task Setup

Unlike typical search problems, the climate researchers need to map every data point in the CAP to the corresponding CDP question/section. We have observed that climate researchers annotate documents in full, similar to how Climate Watch[10] has been constructed. For this reason, it is critical to evaluate the answer-to-question mapping. If we search for answers, then many of the questions will be answered by the summary section of a CAP, which does not help their goals. For example, a CDP question will ask to disclose all climate risks that a corporation may be exposed to. Our aim is to then identify all relevant sections.

Therefore, our procedure is as follows: First, the expert (climate policy researcher on our team and co-author) selected 5 pages at random from a collection of 20 State CAPs and then selected a random paragraph from each page. For this experiment, we use the MSMARCO-MiniLM finetuned on all 3 subsets CDP-CITIES, CDP-STATES and CDP-CORPORATION (our best performing model in the intrinsic evaluation) on each text segment and select the top five scoring CDP-STATES questions. We then presented each segment along with the retrieved questions to the expert and had them annotate the relevance for each question-answer pair on a three point scale: No Relevance, Relevant, Highly Relevant.[11]

6.2 Results

Table 6 shows the climate change researcher's evaluation metrics for our model. Overall, 71.0% of the 500 questions retrieved were judged Relevant and 35.6% rated Highly Relevant. One pitfall of our model is that there were more Highly Relevant predictions ranked fifth than first. One possible explanation for this is that the top retrieved questions were often more general while the questions that were ranked lower were more specific and easier to match (see Table 13 in the Appendix.)

We show some examples in Table 7. The first four examples show high degrees of success. In example 1, our model correctly identifies the state CAP text as impact-related and captures the specific discussion of compound risks. However, example 6 appears to highlight a gap in the CDP questionnaire related to the topic of environmental justice, a result in itself of considerable interest. Although our study is limited, it shows the promise of using our benchmark. Similar extrinsic studies could be conducted on different climate reports such as corporate ESGs or national Paris Agreement plans.

7 Related Work

7.1 Benchmarks in NLP

With the ever growing inventory of deep learning based models on a plethora of NLP tasks, the community is working towards general purpose models that perform well at multiple tasks. Consequently, various benchmarks that evaluate general-purpose understanding have been proposed. For example, GLUE (Wang et al., 2018) and the subsequently more challenging SuperGLUE (Wang et al., 2019) datasets were introduced to test language understanding capabilities. However, doing well on general purpose benchmarks does not translate well to domain-specific tasks like legal reports, biomedical files or numerical problems. This led to the emergence of benchmarks that cater to a particular field of expertise. For instance, LexGLUE (Chalkidis et al., 2022) for legal text and NumGLUE (Mishra et al., 2022) for mathematical reasoning. CLIMABENCH follows on this chain of thought to provide a unified way to evaluate models on CC specific problems.

[10] https://www.climatewatchdata.org/

[11] By construction, in our rating there may be multiple relevant questions found for each text segment.

Ex.	Text Segment from State Climate Action Plans	Top Questions
1	Sea level rise will inundate some nearby coastal areas, and related salt-water intrusion, coupled with increased drought stress may impact water supplies.	Q1: Please describe the current and/or anticipated impacts of climate change. Q3: Please detail any compounding factors that may worsen the impacts of climate change in your region.
2	The afforestation goal is to increase the area of forested lands in the state by 50,000 acres annually through 2025.	Q1: Please provide the details of your region's target(s). Q2: Please provide details of your climate actions in the Land use sector.
3	By a majority vote, the ICCAC presents a policy option that, if deemed necessary, would build one new 1200-megawatt nuclear power plant in Iowa by January 1, 2020.	Q3: Please provide details of your renewable energy or electricity target(s). Q4: Please provide details of your climate actions in the Energy sector.
4	California maintains a GHG inventory that is consistent with IPCC practices ... Reports from facilities and entities that emit more than 25,000 MTCO2e are verified by a CARB-accredited third-party verification body.	Q1: Please give the name of the primary protocol, standard, or methodology you have used to calculate your government's GHG emissions. Q3: Please provide the following information about the emissions verification process.
5	A leading driver of these high emissions is the fact that the District's daytime population swells by 400,000 workers every workday, which is the largest percentage increase in daytime population of any large city in the nation.	Q4: Please indicate if your region-wide emissions have increased, decreased, or stayed the same since your last emissions inventory, and please describe why. Q5: Please report your region-wide base year emissions in the table below.
6	State law defines environmental justice as the fair treatment of people of all races, cultures, and incomes with respect to the development, adoption, implementation, and enforcement of environmental laws, regulations, and policies.	Q1: Please explain why you do not have policies on deforestation and/or forest degradation. Q4: Please provide details of your climate actions in the Governance sector.

Table 7: Examples from our human pilot study in which our climate expert has evaluated the relevance of CDP questions linked to selected text from state climate action plans. A fragment of the matched text is presented with two illustrative questions from the set of five question matches generated by our model.

7.2 Climate NLP

CLIMATEXT (Leippold and Varini, 2020) and CLIMATEFEVER (Leippold and Diggelmann, 2020) extracted and filtered documents from Wikipedia and other sources to curate CC corpus that was further annotated by humans. Due to the challenging and evolving nature of this field, the labelling process is not only time consuming but prone to noise due to poor inter-annotator agreements. Unlike these small annotated datasets, we are able to effectively utilize existing semi-structured disclosure forms for a much larger set of supervised data.

Berrang-Ford et al. (2021) use machine learning to select scientific literature on climate adaptation for the task of systematic evidence mapping. Luccioni et al. (2020) use Task Force on Climate-related Financial Disclosures (TCFD) questionnaires in a QA setup similar to ours while Corringham et al. (2021) use document headers for course-grained classification on Paris Climate Agreements.

8 Limitations and Future Work

One current limitation of our benchmark is that the datasets are English only, thus restricting evaluation to English trained models. Although CDP-STATES has disclosures in other languages it represents a small portion of the reports. We plan to include relevant CC datasets from the multilingual European Union Public Data Catalog[12] in the future, while encouraging contributions from the broader community. Another limitation is that expert human evaluation is missing for our new tasks. We conduct an analysis involving a climate expert for one of our best-performing models on the QA task. But, a general human evaluation of the other classification tasks is not conducted.

We do not thoroughly investigate the efficiency of the Transformer models in this work. But, the compute and training efficiency statistics we provide in A.2 and Table 11 are a step in this direction. We encourage future work on CLIMABENCH to leverage models that are both performant and efficient.

Finally, there exists more types of carbon disclosures (TCFD, SBTi) as well as publicly accessible corporate sustainability reports that we wish to include but require more difficult scraping and data preprocessing.

9 Acknowledgement

This project was funded by the Climate Change AI Innovation Grants program, hosted by Climate Change AI with the support of the Quadrature Climate Foundation, Schmidt Futures, and the Canada Hub of Future Earth.

[12] data.europa.edu

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Lea Berrang-Ford, Anne J Sietsma, Max W. Callaghan, Jan C. Minx, Pauline F. D. Scheelbeek, Neal Robert Haddaway, Andy Haines, and Alan D. Dangour. 2021. Systematic mapping of global research on climate and health: a machine learning review. *The Lancet. Planetary Health*, 5:e514 – e525.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael James Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *ACL*.

Tom Corringham, Daniel Spokoyny, Eric Xiao, Christopher Cha, Colin Lemarchand, Mandeep Syal, Ethan Olson, and Alexander Gershunov. 2021. Bert classification of paris agreement climate action plans. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

Markus Leippold and Thomas Diggelmann. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Markus Leippold and Francesco Saverio Varini. 2020. Climatext: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Bruce Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Fernando Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Alexandra Sasha Luccioni, Emily Baylor, and Nicolas Anton Duchêne. 2020. Analyzing sustainability reports using natural language processing. *ArXiv*, abs/2011.08073.

Prakamya Mishra and Rohan Mittal. 2021. Neuralnere: 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*.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Singh Sachdeva, Peter Clark, Chitta Baral, and A. Kalyan. 2022. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In *ACL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Aarohi Srivastava, Abhinav Rastogi, Abhishek B Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, and Amanda Askell et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022a. 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.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022b. Towards fine-grained classification of climate change related social media text. In *ACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *ArXiv*, abs/2110.12010.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Compute Details

We used a 24 core AMD Ryzen CPU machine with 128 GB RAM for data processing. For training and inference of the deep learning models, we utilize 4 Nvidia RTX 2080Ti GPUs with 11GB memory each. Each model was trained on a single GPU at a time.

A.2 CO2 Emission Related to Experiments

A cumulative of 338 hours of computation was performed on hardware of type RTX 2080 Ti (TDP of 250W). Total emissions are estimated to be 36.5 $kgCO_2eq$. Estimations were conducted using the MachineLearning Impact calculator presented in (Lacoste et al., 2019).

A.3 Pretrained Transformer Models

BERT (Devlin et al., 2019) is a popular Transformer-based language model pre-trained on masked language modeling and next sentence prediction tasks. It makes use of WordPiece tokenization algorithm that breaks a word into several subwords, such that commonly seen subwords can also be represented by the model.

RoBERTa (Liu et al., 2019) uses dynamic masking and eliminates the next sentence prediction pre-training task, while using a larger vocabulary and pre-training on much larger corpora compared to BERT. Another notable difference is the use of byte pair encoding compared to wordPiece in BERT.

DistilRoBERTa (Sanh et al., 2019) leverages knowledge distillation during the pre-training phase reducing the size of the RoBERTa model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. Sanh et al. (2019) originally distilled the BERT model but we utilize the better performing RoBERTa version in our experiments.

Longformer (Beltagy et al., 2020) extends Transformer-based models to support longer sequences with the help of sparse-attention. It uses a combination of local attention and global attention

Section	Category/Label
Hazards: Adaptation	Adaptation
Adaptation	Adaptation
Buildings	Buildings
Hazards: Climate Hazards	Climate Hazards
Hazards: Social Risks	Climate Hazards
Climate Hazards	Climate Hazards
Climate Hazards and Vulnerability	Climate Hazards
Climate Hazards & Vulnerability	Climate Hazards
City-wide Emissions	Emissions
Emissions Reduction	Emissions
GHG Emissions Data	Emissions
Local Government Emissions	Emissions
Emissions Reduction: City-wide	Emissions
City Wide Emissions	Emissions
Emissions Reduction: Local Government	Emissions
Local Government Operations GHG Emissions Data	Emissions
Energy Data	Energy
Energy	Energy
Food	Food
Governance and Data Management	Governance and Data Management
Opportunities	Opportunities
Strategy	Strategy
Urban Planning	Strategy
Transport	Transport
Waste	Waste
Water	Water
Water Security	Water

Table 8: The Sections in the CDP Cities Questionnaire and the corresponding Label assigned by climate expert.

Task	Average	Max	Min	Std
CLIMA-INSURANCE	203	4588	11	326
CLIMA-INSURANCE+	206	4588	11	335
CLIMA-CDP	73	801	11	83
CLIMA-QA	105	834	15	88
CLIMATEXT	23	124	11	10
CLIMATESTANCE	30	98	11	12
CLIMATEENG	30	98	11	12
CLIMATEFEVER	47	311	11	19
SCIDCC	580	2014	13	223

Table 9: Statistics for the number of tokens in each task of CLIMABENCH

mechanism that allows for linear attention complexity and thus makes it feasible to run on longer documents (max 4096 tokens). It however takes much longer to train than the shorter context (512 tokens) models.

SciBERT (Beltagy et al., 2019), a pretrained language model based on BERT, leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks. It was evaluated on tasks like sequence tagging, sentence classification and dependency parsing with datasets from scientific domains. SciBERT gives significant improvements over BERT on these datasets.

ClimateBERT (Webersinke et al., 2021) was warm-started from the DistilRoBERTa model and pretrained on text corpora from climate-related research paper abstracts, corporate and general news and reports from companies that were not publicly released with the model. It was evaluated on tasks like sentiment analysis (using a private dataset), and public datasets like CLIMATEXT and CLIMATEFEVER. In this paper, we evaluate and compare the performance of ClimateBERT on diverse CC tasks for the first time, providing a comprehensive, publicly available and reproducible evaluation.

Model	Source	# Params	Vocab Size	Max Length
BERT	(Devlin et al., 2019)	110M	30K	512
RoBERTa	(Liu et al., 2019)	125M	50K	512
DistilRoBERTa	(Sanh et al., 2019)	82M	50K	512
Longformer	(Beltagy et al., 2020)	149M	50K	4096
SciBERT	(Beltagy et al., 2019)	110M	30K	512
ClimateBERT	(Webersinke et al., 2021)	82M	50K	512

Table 10: Pretrained Transformer Language Models used for Classification tasks

Model	Avg. Runtime (in hours)	Avg. Train Samples/Second	Avg. Train Steps/Second
ClimateBERT	0.40	104.83	1.64
DistilRoBERTa	0.40	101.04	1.58
SciBERT	0.70	53.86	0.84
RoBERTa	0.80	50.46	0.79
BERT	0.85	49.32	0.77
Longformer	14.95	13.82	0.76

Table 11: Compute Efficiency Metrics for the Pretrained Transformer models for the experiments conducted on CLIMABENCH. Models based on the DistilRoBERTa architecture are the most efficient due to smaller model size.

Parameter	Values
loss	hinge, squared_hinge
C	0.01, 0.1, 1, 10
class_weight	none, balanced

Table 12: For the linear SVM, we grid search over the parameters with 5-fold validation to get the best fit out of 80 candidates (16 values * 5 folds) with F1 Macro as the scoring mechanism

Question	MRR@132
Please provide details of your climate actions in the Agriculture sector.	0.870
Please provide details of your climate actions in the Waste sector.	0.789
Please provide details of your climate actions in the Transport sector.	0.774
Please provide details of your climate actions in the Buildings & Lighting sector.	0.597
Please describe these current and/or anticipated impacts of climate change.	0.492
Please complete the table below.	0.487
Please indicate the opportunities and describe how the region is positioning itself to take advantage of them.	0.445
Please provide details of your climate actions in the Energy sector.	0.397
Please describe the adaptation actions you are taking to reduce the vulnerability of your region's citizens, businesses and infrastructure to the impacts of climate change identified in 6.6a.	0.378
Please describe these current and/or future risks due to climate change.	0.327
List any emission reduction, adaptation, water related or resilience projects that you have planned within your region for which you hope to attract financing, and provide details on the estimated costs and status of the project. If your region does not have any relevant projects, please select "No relevant projects" under Project Area.	0.319
Please provide details of your climate actions in the Land use sector.	0.286
Please provide the details of your region-wide base year emissions reduction target(s). You may add rows to provide the details of your sector-specific targets by selecting the relevant sector in the sector field.	0.252
Please describe the adaptation actions you are taking to reduce the vulnerability of your region's citizens, businesses and infrastructure to the risks due to climate change identified in 5.4a.	0.247

Table 13: Question difficulty evaluated on the test set of CDP-STATES ranked from best performing to worst performing. Filtered to only questions that appeared at least twenty times.