

Benchmarking the Benchmarks: Reproducing Climate-Related NLP Tasks

Tom Calamai
Amundi,
Télécom Paris, Inria,
Institut Polytechnique de Paris,
France
tom.calamai@amundi.com

Oana Balalau
Inria,
Institut Polytechnique de Paris,
France
oana.balalau@inria.fr

Fabian M. Suchanek
Télécom Paris,
Institut Polytechnique de Paris,
France
suchanek@telecom-paris.fr

Abstract

Significant efforts have been made in the NLP community to facilitate the automatic analysis of climate-related corpora by tasks such as climate-related topic detection, climate risk classification, question answering over climate topics, and many more. In this work, we perform a reproducibility study on 8 tasks and 29 datasets, testing 6 models. We find that many tasks rely heavily on surface-level keyword patterns rather than deeper semantic or contextual understanding. Moreover, we find that 96% of the datasets contain annotation issues, with 16.6% of the sampled wrong predictions of a zero-shot classifier being actually clear annotation mistakes, and 38.8% being ambiguous examples. These results call into question the reliability of current benchmarks to meaningfully compare models and highlight the need for improved annotation practices. We conclude by outlining actionable recommendations to enhance dataset quality and evaluation robustness.

1 Introduction

As climate change becomes a more and more urgent problem ([The World Meteorological Organization \(WMO\)](#)), there has been a surge of interest in the field of climate-related natural language processing (NLP). Over 60 recent works study tasks such as detecting climate-related discourse, identifying green claims in corporate communications, answering climate-related questions, and detecting deceptive patterns in climate-related communication ([Calamai et al., 2025](#)). And yet, the progress in this area is hard to gauge, for several reasons: First, the approaches are evaluated on different datasets even if they treat the same tasks; second, the evaluations differ in their metrics, which makes the results incomparable; third, the approaches are rarely compared to simple TF-IDF baselines, so that it is unclear to what degree the proposed problems can be solved by simpler methods as well; and finally,

few of the works use zero-shot LLMs as competitors, so that it is hard to judge to what degree the proposed methods have become obsolete with the arrival of generative models.

In this paper, **we set out to make a transversal and unified comparison of climate-related NLP tasks**. We compiled 29 climate-related NLP datasets and evaluated a suite of baselines that cover traditional machine learning approaches such as TF-IDF with logistic regression, as well as finetuned Transformer models, and open and closed source LLMs in a standardized setting – effectively building a benchmark. We also performed an in-depth error analysis, with more than 500 manually annotated errors.

Our work yields the following insights:

- (1) TF-IDF, finetuned models, and zero-shot LLMs perform similarly, with finetuned models performing slightly better. While this confirms that current models are capable of solving these tasks, it also questions the difficulty of the tasks. If TF-IDF performs well, as we systematically show, then this means that the datasets consist of simple examples that can be predicted using word-frequency only.
- (2) 96% of our datasets contain annotation issues, with 16.6% of the sampled wrong predictions of a zero-shot classifier being actually clear annotation mistakes, and 38.8% being ambiguous examples where multiple labels could fit or where it is not clear which label to assign.
- (3) Our manual analysis shows that finetuned models may perform better not because they generalize better, but because they overfit on annotation noise, bias, or unclear guidelines.

These findings highlight the limitations of the datasets to meaningfully compare models, because their simplicity and inconsistencies compress performance into a narrow band, masking real differences in model capability.

We believe our findings motivate the need for

more challenging datasets, with examples that go beyond keywords, require reasoning, and come with clearer, more robust annotation schemes. We propose recommendations for construction such datasets at the end of our paper.

We make all cleaned-up datasets, baselines, and a Python library to run comparisons publicly available at https://github.com/tcalamai/acl_climateNLPtoolbox.

2 Related Work

Climate-related Benchmarks. To the best of our knowledge, only three benchmarks have been proposed for climate-related tasks: **ClimaBench** (Spokoyny et al., 2023) contains the existing datasets climateStance, climatext, climateEng, climateFEVER, SciDCC, and introduces 3 additional datasets, ClimaINS, ClimaQA, and ClimaTOPIC. The authors evaluate multiple fine-tuned transformers and simpler baselines on ClimaBench. **ClimateGPT** (Thulke et al., 2024) is a benchmark that consists of ClimaBench, Pira 2.0 MCQ (Pirozelli et al., 2023) and CC-Contrarian Claims. Several LLMs were evaluated in a few-shot setting. Finally, **ClimRetrieve** (Schimanski et al., 2024b) is a benchmark restricted to climate-related information retrieval. We improve upon these efforts as follows:

1. Our dataset collection is significantly larger than previous ones, with 19 more datasets than ClimaBench and 17 more datasets than ClimateGPT
2. We evaluate simple baselines, fine-tuned transformers, and recent LLMs in zero-shot settings, thereby contextualizing all performances
3. We conduct an in-depth analysis of the errors

Model Variability and Annotation Errors. Reproducibility and significance of results is a major issue in the scientific literature, especially in the machine learning community (Bouthillier et al., 2021; Gundersen et al., 2022; Ruffinelli et al., 2020). One challenge is that not all papers share the code, the experimental settings, the datasets created and the dataset annotation details. This issue has recently gained a lot of attention in the community, and conferences guidelines now often explicitly ask authors for these elements. Another source of irreproducibility comes from the machine learning algorithms themselves, and the design choices that can be made. Gundersen et al. (2022) identified 41 design choices that can affect reproducibly,

among which: hyperparameter tuning that is costly and is often done manually (Bouthillier and Varoquaux, 2020); dataset issues such as preprocessing, data splits and annotation quality; a lack of sufficient baselines; and a lack of confidence intervals for results. Annotation quality is another major issue in the machine learning community, and solutions such as automatic detection of erroneous labels (Klie et al., 2023a) and datasheets for datasets (Gebru et al., 2021) have been proposed. In this work, we apply these considerations to climate-related NLP tasks.

3 Tasks, Datasets, and Models

3.1 Tasks

In previous work (Calamai et al., 2025), we have identified the following climate-related NLP tasks:

Task 1. Climate-Related Topic Detection: Given an input sentence or a paragraph, output a binary label, “climate-related” or “not climate-related”.

Task 2. Thematic Analysis: Given an input sentence or a paragraph, output a subtopic related to climate change. The categories can be the four categories of the Task Force on Climate-related Financial Disclosures (TCFD), the categories of Environment, Social and Governance (ESG), or custom-made categories.

Task 3. Climate Risk Classification: Given an input sentence or a paragraph, output the label “opportunity” (if the input talks about a positive effect of climate-change for the company) or “risk” (if the input talks about a negative effect). Some works focus only on risks, classifying them into types of risks (e.g., physical risk, reputational risk, regulatory risk, or transition risk).

Task 4. Green Claim Detection: Given an input sentence or a paragraph, output a binary label, “green claim” or “not green claim”. Green claims refer to the practice of suggesting or otherwise creating the impression that a product or a service is environmentally friendly (Stammach et al., 2023).

Task 5. Green Claim Characteristics: Given an input sentence or a paragraph labeled as a green claim, output a more fine-grained characterization of the claim. This is a multi-label classification task; the labels can be about the form (e.g. specificity) or the substance (e.g. action, targets, facts).

Task 6. Green Stance Detection: Given two input sentences or paragraphs, one labeled as the claim and one as the evidence, predict if the evidence

supports the claim, refutes the claim, or is neutral towards the claim. Some studies fix the claim (e.g. the claim is always “Climate change poses a severe threat”) and aim to predict if the evidence supports or refutes that claim.

Task 7. Climate-Related Question Answering: Given an input question and a set of resources (paragraphs or documents), produce an answer to the question.

Task 8. Classification of Deceptive Techniques: Given a statement, classify it into argumentative categories, such as fallacies, types of arguments, or rhetorical techniques.

3.2 Datasets

For each of the tasks, our previous work (Calamai et al., 2025) listed all related datasets. For this work, We collected all those datasets that are openly available to reproduce the original results. Table 1 shows all these datasets with their description from Calamai et al. (2025). These datasets suffered from several issues, and hence we subjected them to a data cleaning pipeline. We provide here an overview of the issues we encountered, and we give the detailed statistics of each cleaning step in Appendix A.

Duplicate Removal. We identified many duplicates in the datasets, some of which became visible only after correcting formatting issues such as differing numbers of spaces. We differentiate between exact duplicates (same text and same label) and those with conflicting labels. Exact duplicates were removed to keep only one instance. For duplicates with conflicting labels, when possible we reconstructed the dataset to avoid them (e.g., Spokoiny et al. (2023) provide the source questionnaires for *ClimaQA* and *ClimaINS*); otherwise, we removed them. Some duplicates were intrinsic to the task, in which case we did not remove them (e.g. *ClimateFEVER evidence* contains pairs of claim-evidence; the same evidence could be found for multiple claims).

Noisy text. We investigated the text quality of datasets using a gibberish detection model¹ and a language detection model². We found that some datasets have noisy text samples or non-English

text, but in small proportion (<0.5%). Moreover, noisy text is usually labeled with the negative label. For example, for detecting “climate-related” vs “not climate-related”, the noisy text had the label “not climate-related”. The few positively labeled noisy texts were mostly false positives – they were actually not noisy. Finally, even noisy examples are part of real-world texts. We therefore did not remove noisy samples.

Input text length. BERT-like models usually have a limited context window (e.g. 4096 tokens for Longformer). In multiple datasets, we found some excessively long texts (text with more than 4000 tokens), often due to formatting issues (e.g. spaces between all characters). To address this, we applied a text cleaning step to remove formatting and encoding issues using the clean-text³ Python library. Some datasets still contained very long texts, so we removed text longer than 4000 tokens to fit within Longformer’s context window. This impacted only 6 datasets and removed less than 0.3% of the data per dataset. We also identified very short texts resulting from parsing errors in PDF documents. These were typically page numbers, escape sequences, or table fragments. To remove these spurious occurrences, we removed all texts with fewer than 5 tokens.

Dataset size. After data cleaning, 7 of our 29 datasets still contained a large number of samples, ranging from 500 to over 150K. To reduce computational costs and experiment duration, we limited the training and development splits to 10,000 samples, while keeping **the original size of the test splits intact**. During down-sampling, we ensured that label distribution was preserved through stratification. For heavily imbalanced datasets, we adjusted the sampling to improve the balance.

Dataset splits. We split each dataset into 80% train, 10% test and 10% development. We kept the original splits when they were available in these proportions. However, some existing splits exhibited train-test contamination, in which case we re-created the splits.

3.3 Models and Measures

Models. We evaluated a range of models on each dataset: random baselines, a traditional approach (TF-IDF with logistic regression), fine-tuned transformers (distilRoBERTa, Longformer), and large

¹<https://huggingface.co/madhurjindal/autonlp-Gibberish-Detector-492513457>

²<https://huggingface.co/papluca/xlm-roberta-base-language-detection>

³<https://github.com/prasanthg3/cleantext>

Dataset	Input	Labels
Climate-Related Topic Detection		
ClimateBug-data, Yu et al. (2024)	sentences from Banks' reports	<i>relevant/irrelevant</i> : Climate change and sustainability (including ESG, SDGs related to the environment, recycling and more)
ClimateBERT's climate detection, Bingler et al. (2023)	paragraphs from reports	<i>1/0</i> : Climate policy, climate change or an environmental topic
ClimateText (Wikipedia, 10-K, claims), Varini et al. (2020)	sentences from Wikipedia, 10-Ks or web scraping	<i>1/0</i> : Directly related to climate-change
ClimateText (wiki-doc), Varini et al. (2020)	sentences from a Wikipedia page	<i>1/0</i> : Extracted from a Wikipedia page related to climate-change
Sustainable signals's reviews, Lin et al. (2023)	online product reviews (user comments)	<i>relevant/irrelevant</i> : Contains terms related to sustainability
Thematic Analysis		
TCFD rec., Bingler et al. (2021)	paragraphs from corporate annual reports	<i>Metrics and Targets, Risk Management, Strategy, Governance and General</i> : TCFD 4 main categories
ESGBERT's ESG, Schimanski et al. (2023b)	Sentences from reports and corporate news	<i>Environment, Social, Governance and None</i> : Environmental criteria comprise a company's energy use, waste management, pollution, as well as compliance with governmental regulations. Special areas of interest are climate change and environmental sustainability.
ESGBERT's Nature, Schimanski et al. (2024a)	Paragraphs from reports	<i>General, Nature, Biodiversity, Forest, Water</i> : Multi-label Nature-related topics
SciDCC, Mishra and Mittal (2021)	News articles (Title, Summary, Body)	<i>Environment, Geology, Animals, Ozone Layer, Climate, etc.</i> : Category in which the article was published (Automatic Label)
ClimateEng, Vaid et al. (2022)	Tweets posted during COP25 filtered by keywords (relevant to climate-change)	<i>Ocean/Water, Politics, Disaster, Agriculture/Forestry, General</i> : Sub-categories of climate-change
ClimaTOPIC, Spokoyny et al. (2023)	CDP responses (short texts)	<i>Adaptation, Buildings, Climate Hazards, Emissions, Water, etc.</i> : Category of the question (Automatic Label)
Climate Risk Classification		
ClimateBERT's Sentiment, Bingler et al. (2023)	Paragraphs from companies' annual reports	<i>Risk, Opportunity, Neutral</i> : <i>Risk</i> or threat that negatively impacts an entity of interest (negative sentiment); or <i>Opportunity</i> arising due to climate change (positive sentiment); <i>Neutral</i> otherwise.
Green Claim Detection		
Green Claims, Woloszyn et al. (2022)	Marketing Tweets	<i>Green Claim/Not Green</i> : Environmental (or green) advertisements refer to all appeals that include ecological, environmental sustainability, or nature-friendly messages that target the needs and desires of environmentally concerned stakeholders.
Environmental Claims, Stammbach et al. (2023)	Paragraph from reports	<i>Yes/No</i> : Environmental claims refer to the practice of suggesting or otherwise creating the impression [...] that a product or a service is environmentally friendly (i.e., it has a positive impact on the environment) or is less damaging to the environment than competing goods or services [...] In our case, claims relate to products, services, or specific corporate environmental performance.
Green Claim Characteristics		
Implicit/Explicit Green Claims, Woloszyn et al. (2022)	Marketing Tweets	<i>Implicit green claims</i> raise the same ecological and environmental concerns as <i>explicit green claims</i> (see definition in Section C.6), but without showing any commitment from the company. If the tweet does not contain a green claim then <i>No Claim</i> .
Specificity, Bingler et al. (2023)	Paragraph from reports	<i>Specific, Non-specific</i> : A paragraph is <i>Specific</i> if it includes clear, tangible, and firm-specific details about events, goals, actions, or explanations that directly impact or clarify the firm's operations, strategy, or objectives. <i>Non-specific</i> otherwise.
Commitments and Actions, Bingler et al. (2023)	Paragraph from reports	<i>Yes/No</i> : A paragraph is a commitment or an action if it contains targets for the future or actions already taken in the past.
Net Zero/Reduction, Schimanski et al. (2023a)	Paragraph from Net Zero Tracker, Lang et al. (2023)	<i>Net-zero, Reduction, None</i> : The paragraph contains either a <i>Net-Zero</i> target, a <i>Reduction</i> target or no target (<i>None</i>)
Green Stance Detection		
ClimateFEVER (evidence), Diggelmann et al. (2020)	A claim and an evidence sentence from Wikipedia	<i>Support, Refutes, Not Enough Information</i> : Determines the relation between a claim and a single evidence sentence
LobbyMap (Stance), Morio and Manning (2023)	Page from a company communications (report, press release, ...)	<i>Strongly supporting, Supporting, No or mixed position, Not supporting, Opposing</i> : Given the policy and the page, classifies the stance
Global Warming Stance Detection (GWSD), Luo et al. (2020)	Sentences from news about global warming	Stance of the evidence (<i>Agree, Disagree, Neutral</i>) toward the claim: Climate-Change is a serious concern.
ClimateStance, Vaid et al. (2022)	Tweets posted during COP25 filtered by keywords (relevant to climate-change)	Stance towards climate change prevention: <i>Favor, Against, Ambiguous</i> . (Stance used as a broad notion including sentiment, evaluation, appraisal, ...)
ClimateFEVER (claim), Diggelmann et al. (2020)	A claim and multiple evidence sentences from Wikipedia	<i>Support, Refutes, Debated, Not Enough Information</i> : Determines if a claim is supported by a set of retrieved evidence sentences
LobbyMap (Page), Morio and Manning (2023)	Page from a company communications (report, press release, ...)	<i>1/0</i> : Contains a stance on a remediation policy
LobbyMap (Query), Morio and Manning (2023)		<i>GHG emission regulation, Renewable energy, Carbon tax, ...</i> : Classifies the remediation policy
Climate-Related Question Answering		
ClimaQA, Spokoyny et al. (2023)	The text from a <i>response</i> to one of the CDP questions; and one of the <i>questions</i> from the CDP questionnaire	<i>1</i> : the response answers this question <i>0</i> : The response does not answer this question, but another one
ClimaINS, Spokoyny et al. (2023)	The text from a <i>response</i> to one of the questions from the NAIC questionnaire	<i>MANAGE, RISK PLAN, MITIGATE, ENGAGE, ASSESS, RISKS</i> : The labels correspond the 8 questions asked in the NAIC questionnaires
Classification of Deceptive Techniques		
LogicClimate, Jin et al. (2022)	texts from climatefeedback.org	<i>Faulty Generalization, Ad Hominem, Ad Populum, False Causality, etc.</i> : Classifies fallacies (Multi-label)
Contrarian Claims, Coan et al. (2021)	paragraphs from conservative think tank	<i>No Claim, Global Warming is not happening, Climate Solutions won't work, Climate impacts are not bad, etc.</i> : Classifies arguments into super/sub-categories of climate science denier's arguments

Table 1: Description of the datasets we collected.

language models (LLMs) in a zero-shot setting (GPT-4o-mini, Llama 3.1 – 8B and 70B).

Evaluation. All models are evaluated on the test datasets. We computed the macro F1-score for each dataset, even if the original papers may have reported other scores such as accuracy or the micro F1-score.

Statistical Significance. To compare models, we computed a 95% confidence interval (CI) of the F1-scores based on bootstrapping (Efron and Tibshirani, 1994). This interval measures the variability based on the data sampling. We consider a difference *significant* if the CIs are disjoint.

Parameters. Throughout our experiments, we used fixed parameters following Spokoyny et al. (2023)’s parameter choices and seeds to control for randomness (detailed parameters in Appendix B).

4 Results and Analysis

4.1 Results

Table 2 shows the macro-F1 of all models on all datasets. We also show the most relevant performance score from the original studies, which is not necessarily the macro F1-score and thus not comparable to our results (see details in the table caption). We observe that the finetuned Transformer models generally perform best, with an average F1-score of 74.73% for distilRoBERTa and 74.04% for Longformer. Most datasets (70%) do not contain long text, which limits the benefit of using Longformer compared to distilRoBERTa. Secondly, the zero-shot LLMs performed competitively, with an average F1-score reaching 69.78% for GPT-4o-mini. As expected, the larger Llama 3.1 70B outperformed Llama 3.1 8B, although not by a large margin. Finally, the TF-IDF baseline performed surprisingly well, with an average F1-score of 69.18%, out-performing the Llama models, and coming close to GPT-4o-mini. This indicates that the tasks studied in these datasets can be largely solved using only term-frequency features.

4.2 Error Analysis

It is surprising that the top performing LLMs perform worse than the fine-tuned models, despite being significantly more complex and having demonstrated high performance on many other benchmarks (Meta, 2024). We hypothesize that a significant portion of the model’s errors might in fact be

issues in the datasets. To test this hypothesis, we analyzed a sample of 503 errors. We investigated the errors of GPT-4o-mini, the best-performing LLM, to minimize the number of genuine errors. However, our findings are relevant for the other LLMs as well, as we found that 70% of the sampled errors are also errors made by the Llama’s models, and 70% of Llama’s errors are also GPT-4o-mini errors (Table 7 in Appendix C).

We analyzed the discrepancies between the model’s prediction and the gold standard to understand both model and dataset limitations. We distinguish mislabeled from debatable instances to align with existing frameworks that distinguish between annotation errors from Human Label Variation (Weber-Genzel et al., 2024).

We sampled between 20 and 30 errors per dataset. Some datasets had fewer than 20 errors, in which case we reviewed them all. Given the input text, the true label, and the prediction of the model, we classified them based in type of error.

Model errors. Incorrect predictions made by the zero-shot GPT-4o-mini approach account for 44.6% of the sampled errors – these are clear misclassifications, not cases of ambiguity or errors in the gold standard. They are caused by different factors such as hallucinations, misunderstanding of implicit or indirect relations, anchoring to a specific word in the prompt, not following part of the instructions, or having no identifiable cause. Many of these errors could most likely be fixed by using a larger model, and will not be detailed here (see Appendix C.11 for details).

Annotation mistakes. Evaluation datasets can contain annotation errors that bias the performances and the conclusions than can be drawn from experiments (Weber-Genzel et al., 2024; Rucker and Akbik, 2023; Bowman and Dahl, 2021; Klie et al., 2023b). We found obvious annotation errors in 60% of the datasets. This accounts for 16.6% of the sampled errors.

Debatable Errors. Finally, we find debatable examples in almost all datasets (96%). We identified three main causes: statements that are out-of-context; statements that are ambiguous or implicit; and definitions of labels that are not exhaustive or detailed enough, leaving room for interpretation. 38.8% of the sampled errors are debatable.

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama 8B	Llama 70B	Reference
Climate Topic Classification								
✓ climateBUG-data	49.6±0.6	86.4±0.4	90.5±0.4	90.5±0.3	89.2±0.4	79.0±0.5	88.3±0.3	91.36 ^f
✓ ClimateBERT Cl. det.	42.8±4.9	79.3±4.8	95.8±2.4	94.0±3.0	93.2±3.1	88.3±3.8	93.4±3.2	99.1 ^{3,c}
▲ climatext (Wiki-doc)	49.3±1.5	80.4±1.2	85.6±1.1	83.8±1.2	84.0±1.2	86.3±1.1	89.8±1.0	-
✓ climatext (10k)	46.9±5.4	91.0±4.0	97.0±2.4	96.5±2.6	95.4±3.2	90.9±4.2	96.0±2.8	95±2 ^{1,t}
✓ climatext (claim)	48.2±3.1	74.7±2.8	67.8±3.2	75.4±2.9	82.2±2.4	83.6±2.2	85.7±2.0	83±1 ^{1,t}
✓ climatext (Wikipedia)	40.1±4.8	83.5±7.7	83.9±7.2	88.2±6.5	84.7±7.2	82.8±7.6	87.4±6.5	80±6 ^{1,t}
▲ Sustainable Signals rev.	37.6±11.0	65.8±11.1	73.2±10.0	73.7±9.9	69.7±11.0	63.5±10.7	73.3±9.9	-
Climate Sub-thematic topic classification								
▲ ClimaTOPIC	6.0±0.5	46.6±1.2	55.8±1.4	54.8±1.3	35.8±1.0	34.5±1.0	30.5±1.0	65.22 ^r
✓ climateEng	13.2±2.9	58.4±9.3	70.5±8.5	67.3±8.6	65.8±9.2	53.9±9.4	61.1±8.5	74.58 ^r
✓ esgbert Biodiversity	41.1±6.4	91.3±5.5	91.3±5.1	89.0±5.5	91.8±4.9	88.8±5.6	82.3±5.9	92.29 ^{1,g}
✓ esgbert Forest	41.4±5.9	92.9±4.8	97.2±2.8	97.0±2.8	91.6±4.9	87.4±5.6	86.2±5.8	95.37 ^{1,g}
✓ esgbert Nature	51.5±6.5	82.9±5.0	89.8±3.9	89.0±4.0	82.9±5.2	86.7±4.5	88.0±4.6	94.19 ^{1,g}
✓ esgbert Water	42.7±6.4	84.6±6.6	95.3±3.4	93.9±3.9	93.8±4.1	89.4±5.3	93.5±3.9	95.10 ^{1,c}
▲ sciDCC	3.7±1.0	42.6±3.9	39.8±2.8	42.1±3.3	29.2±2.3	20.5±2.2	30.0±2.4	54.79 ^l
✓ Climate TCFD rec.	16.5±3.7	56.5±5.6	69.4±5.2	68.5±5.2	47.5±5.4	43.2±5.2	50.2±5.4	-
✓ ESGBERT E	42.6±7.1	88.1±5.1	95.8±3.1	95.8±3.2	95.0±3.5	95.1±3.5	94.6±3.4	93.19 ^{1,e}
✓ ESGBERT G	53.5±7.1	80.2±6.3	78.2±6.4	83.7±6.4	83.5±5.7	75.6±6.1	62.6±6.9	78.86 ^{1,e}
✓ ESGBERT S	54.4±6.9	82.6±5.6	88.3±4.4	89.6±4.2	78.1±5.8	77.9±6.0	73.6±6.2	91.90 ^{1,e}
Sentiment Analysis								
✓ climate sentiment	35.9±5.2	69.0±5.5	79.9±4.8	77.7±4.8	77.3±4.9	71.4±5.3	70.5±5.4	83.8 ^{3,c}
Claim detection								
✓ Environmental Claims	42.8±5.5	80.5±5.4	90.5±3.8	91.2±3.8	86.8±4.5	81.4±5.0	76.6±5.4	84.9 ^{1,r}
▲ Green Claims	46.1±10.9	86.1±8.0	94.5±5.0	97.2±3.5	91.5±6.8	90.0±7.3	86.8±7.3	92.08 ^{1,r}
Claim characteristics								
✓ Commitments&Actions	48.5±5.5	72.7±5.4	76.7±5.1	81.9±4.7	67.2±5.4	64.6±5.2	50.7±5.6	81 ^{3,c}
✓ Climate Specificity	48.9±5.3	72.4±5.0	77.3±4.8	77.5±5.1	71.6±4.9	76.5±4.6	68.0±5.2	77 ^{3,c}
▲ ESGBERT action500	44.4±14.3	82.5±11.0	85.9±10.2	89.1±8.8	76.0±12.6	67.0±13.2	63.0±14.2	-
▲ Implicit/Explicit Claims	37.2±9.9	70.3±12.8	63.0±13.1	81.2±11.9	81.6±9.6	68.5±13.2	75.7±11.3	81.45 ^{2,r}
✓ Net-Zero/Reduction	29.8±4.9	94.8±2.4	97.8±1.5	97.7±1.8	97.3±1.8	93.7±2.8	95.6±2.2	98.7 ^{1,c}
Stance classification								
✓ climateF. claim (our split)	15.4±5.1	35.0±8.3	31.9±7.1	32.5±7.4	19.3±5.9	30.0±7.7	26.3±6.5	80.7 ^{t,t}
✓ climateF. claim (agg.)	-	-	42.5±7.8	41.3±8.6	50.4±8.1	-	-	60.10±8.6 ^h
✓ climateFEVER evidence	28.4±3.1	46.2±4.0	52.7±4.1	51.3±3.9	60.0±3.7	52.5±3.7	63.7±3.9	68.03±12.9 ^h
✓ climateStance	21.3±3.7	49.6±6.5	56.6±6.8	56.1±7.7	56.4±5.9	48.5±5.9	58.0±6.3	59.69 ^r
✓ Global-Warming Stance	29.5±6.5	59.2±7.4	69.3±6.8	75.3±6.5	69.8±6.6	64.4±6.8	72.5±6.5	73 ^t
▲ LobbyMap (Pages)	46.3±0.8	73.1±0.8	71.3±0.8	73.6±0.8	63.2±0.8	62.5±0.8	60.5±0.8	-
▲ LobbyMap (Query)	12.4±0.5	49.3±1.8	57.3±2.1	52.1±2.1	36.4±1.8	27.6±1.6	32.7±1.6	-
▲ LobbyMap (Stance)	19.2±1.3	43.6±1.8	46.7±1.6	44.7±1.9	30.0±1.6	26.9±1.5	24.1±1.5	-
Question answering								
▲ ClimaNS (our split)	12.8±1.8	81.2±2.1	77.6±4.2	75.8±4.5	58.5±2.6	45.4±2.5	48.4±2.6	-
▲ climaQA (our split)	50.5±1.1	50.3±1.0	89.5±0.6	89.4±0.6	78.8±0.8	57.4±1.1	76.1±1.0	-
Deceptive technics								
✓ CC-Contrarian Claims	3.2±0.7	62.7±2.5	71.6±2.7	71.2±2.5	58.8±2.9	40.2±3.1	55.8±2.9	79 ^j
▲ logicClimate	13.3±2.0	13.3±5.0	15.6±5.6	9.4±3.2	27.4±7.8	10.3±4.3	22.1±6.4	29.37 ^k
Average	34.24	69.18	74.04	74.73	69.78	65.03	67.39	-

Table 2: Macro F1 score of baselines and methods on each dataset. The bootstrap confidence interval is displayed as $\pm \frac{F1_{95\%} - F1_{5\%}}{2}$. For each Reference value, we add a clarifying number if the metric is not macro F1: 1. binary F1-score, 2. average F1-score (macro/micro is not specified), 3. weighted average F1-score. In addition, we specify the method in the reference via letters: *a.* SVM, *c.* climateBERT, *d.* Contaminated split, SciBERT, *e.* ESGBERT, *f.* climateBUG-LM, *g.* EnvironmentalBERT, *h.* Human Baseline using annotations, *i.* Filtered Split (by removing disputed claims), *j.* RoBERTa + Logistic Regression, *k.* Electra, *l.* Longformer, *r.* RoBERTa, *t.* BERT. ▲: weakly labeled datasets, ▲: small datasets (<1000 in training dataset), ✓ larger, human-annotated datasets. In red: performance not significantly higher than random; In gray: performance not significantly higher than TF-IDF baseline; In green: fine-tuned models performing significantly better than all zero-shot approaches (and inversely). A performance is considered significantly higher if 0 is not included in the bootstrap 95% confidence interval of the difference of F1-score.

4.3 Dataset Issues

In the following, we discuss possible causes of the annotation mistakes we observed in our study (This section contains a condensed analysis of errors, we detail error per dataset in Appendix C).

Ambiguous annotations guidelines. For each of the datasets, we designed the prompts based on the annotation guidelines. However, we observed that many prediction errors from GPT-4o-mini were caused by ambiguities in the guidelines:

- (1) In ClimateBUG, the guideline says that “Sustainability not related to the environment (i.e. sustainable profits)” should be classified as “Not Climate”. However, some sentences do not specify whether the sustainability is related to the environment or not (e.g. “They are the source of jobs, innovation, sustainability, and prosperity”).
- (2) In ESGBERT Nature, it is not clear if the “water” label, which focuses on “water management, consumption, and pollution”, includes water-related natural disasters such as tsunamis.
- (3) In climate sentiment, it is not clear if corporate ambitions should be classified as “Opportunities” as they “associate specific positive adjectives to the anticipated, past, or present developments and topics covered” (e.g. “TD recently launched a bold and ambitious climate action plan to address the challenges of climate change. This includes a target to achieve net-zero greenhouse gas emissions in our operations and financing activities by 2050”) or as a neutral statement as aligning with climate objectives will not necessary bring growth to a company.
- (4) Environmental Claims does not provide instructions on implicit claims.
- (5) Green Claims is not specific on how to annotate claims suggesting that a product is better, healthier, or has good properties thanks to a natural ingredient.
- (6) In Commitments&Actions, it is not clear how to annotate descriptions of company values, governance structures, or descriptions of existing processes.
- (7) Climate Specificity does not describe how to handle the granularity of the specificity (e.g. “For our sustainable strategy range, we incorporate a series of proprietary ‘red lines’ in

order to ensure the poorest-performing companies from an ESG perspective are not eligible for investment”).

- (8) In ClimateFEVER evidence, it is not clear if “SUPPORTING” means that the evidence is sufficient to entail the claim, or if it means that the evidence is merely in line with the claim.
 - (9) In ClimateStance, it is not clear how to annotate a tweet that is in “opposition to climate change policies”, but simultaneously acknowledges the urgency of climate change.
 - (10) In CC-Contrarian Claims, some labels are difficult to differentiate such as “Weather is cold/snowing” and “Ice/permafrost/snow cover isn’t melting”.
 - (11) In LogicClimate, the “intentional” label is broad and encompasses all other fallacy types.
- These issues impact many datasets and a large part of the sampled errors of GPT-4o-mini (4/9 on ESGBERT Water, 5/20 on ClimateBUG, 5/10 on climate sentiment, 5/14 on Environmental Claims, 5/6 on (Woloszyn et al., 2022), 9/20 Commitments and Actions, 4/20 on specificity, 20/20 additional samples⁴ ClimateFEVER evidence, 2/10 in ClimateStance, 2/10 in CC-Contrarian Claims, 3/10 in LogicClimate).

Implicitness. Many errors were due to the implicit nature of the text. This represents a large portion of the sampled errors in multiple datasets (4/17 on Climate detection, 3/10 on climate TCFD recommendations, 2/8 on ESGBERT, 3/14 on Environmental claims, 2/10 on ClimateStance, 7/10 on GWSD). Interestingly, we even found examples where the model understood the indirect link but not the annotators (2/10 on climate sentiment). A notable implied reference to climate change, which the model struggles on, is categorizing statements about energy as climate/environment-related (4/17 Climate detection, 2/8 on ESGBERT, 3/14 on Climatext (wiki)). These cases are prediction errors by GPT-4o-mini; however, due to the implicit nature of the statements, the labels are debatable.

Multi-label. In many different contexts, multiple labels might be valid (Weber-Genzel et al., 2024). This is what we observed in 26% of the single-label datasets, and 8.75% of the sampled errors. For example, “Reduce GHG emissions [...] by 55%

⁴We randomly sampled 10 errors of GPT-4o-mini predicting “REFUTES” instead of “NOT ENOUGH INFO” and 10 errors of it predicting “NOT ENOUGH INFO” instead of “SUPPORTS”.

by FY2030 and to zero by FY2050 [...]” contains both a reduction target and a net-zero target. We believe this issue arises from trying to simplify the annotation process and build a single-label dataset in settings that are intrinsically multi-label.

Weak labels. Zero-shot LLMs performed significantly worse than finetuned models on weakly labeled datasets that were not annotated through a controlled annotation procedure. The weak labels are useful to create large datasets without requiring human annotations; however, they lack the rigorous definitions that manual labels provide, which are derived from strict annotation guidelines. This automatic generation can introduce inconsistencies such as evolving label definitions over time or unexpected responses in survey data. This makes these labels noisy and less reliable for classification tasks. This explains the lower performances on SciDCC, ClimaINS, ClimaTOPIC and ClimaQA. Additionally, since the labels do not follow strict guidelines, designing a prompt for these tasks requires guessing the actual label meaning, which might be different from the label intended purpose. For example, the Climatext task is to identify climate-related sentences. However, to create the datasets annotated by humans, [Varini et al. \(2020\)](#) relied on a weakly labeled dataset to filter potentially climate-related sentences. Therefore, the intended purpose is climate-relatedness, but the actual meaning of the label is “a sentence extracted from a climate-related Wikipedia page” (which is the prompt used to reach F1-scores above 80%). Despite these non-intuitive definitions, the labels are actually linguistically distinguishable, as shown by the performances of the fine-tuned models and the TF-IDF baselines.

Exhaustivity. This is a particular case of issues arising from weakly labeled datasets. Because they do not rely on rigorous annotation guidelines intended for classification, the annotations are not necessarily exhaustive. The LobbyMap dataset, e.g., is constructed using data from [LobbyMap.org](#), a website that identifies companies’ stances on climate change mitigation policies and quotes companies’ documents with evidence about their stance. Therefore, for each policy, they identify evidence, but they do not necessarily collect all the evidence about that stance. Inversely, the model is asked to identify, given a document, what is the list of policy stances mentioned. GPT-4o-mini tends to predict more stances per document than the original labels. Moreover, in the sampled errors, we found that the

predictions were mostly reasonable. This suggests that the annotations might not be exhaustive.

4.4 Discussion

The weakly labeled datasets are interesting as they build a foundation for relevant tasks: e.g. LobbyMap for identifying stance on climate-remediation policies, SciDCC to classify news, or ClimaQA and ClimaINS—from **Task 7**— and ClimaTOPIC to structure company documents; however, the poor performance and the error analysis reveal that the datasets need further annotations.

Task 1 to 5 focus on topic classification, risk and green claim. When excluding weakly labeled datasets, performances are high for fine-tuned models, but also for TF-IDF, showing that these tasks are highly based on vocabulary. Some datasets remain challenging: TCFD recommendations, ClimateEng and, to a lesser extent, climate sentiment and specificity. However, improving the guidelines to resolve ambiguous cases would improve label consistency, and most likely performances, particularly for LLMs which rely more heavily on guidelines, while finetuned models rely on the annotations.

Task 6 focuses on stance detection. Models demonstrate that they have predictive power, but the performance is relatively low (F1-score under 75%). However, the task is inherently hard, as shown by the low performance of humans on ClimateFEVER, or moderate IAA (Krippendorff’s $\alpha = 0.54-0.64$) reported by [Luo et al. \(2020\)](#).

Task 8 focuses on deceptive techniques. On CC-contrarian claims, given the high number of classes, the performances are high, even for TF-IDF. For fallacy detection, the task is inherently subjective ([Helwe et al., 2024](#)). This is particularly the case when using real-world examples where multiple types of fallacies could fit.

4.5 Recommendations

As shown through our error analysis, many false predictions of a zero-shot model are actually mislabeled or debatable examples. Such examples are problematic for two reasons: First, the performance differences between models are often minimal (sometimes less than 1%), making it difficult to draw meaningful comparisons when the datasets are noisy. Second, it becomes unclear whether observed performance gains stem from genuine model improvements or from overfitting to annotation biases and errors. Fine-tuned models may perform well on these datasets, but that does not mean

that the datasets are clean: The models may pick up patterns in the data that have been informally discussed between the annotators, but that are not codified in the guidelines. To address these issues and improve the reliability of model evaluation, we propose the following set of recommendations.

Use Precise and Exhaustive Annotation Guidelines. Ambiguous or under-specified guidelines result in inconsistent annotations, reducing dataset reliability and skewing model evaluation. To mitigate this, guidelines should be:

- (1) Precise: Clearly define all terms and concepts.
- (2) Exhaustive: Address edge cases and ambiguities.
- (3) Unambiguous: Avoid vague phrasing and provide concrete examples.

Additionally, we recommend not using automatically labeled datasets, as they do not follow well-defined annotation procedures and can easily produce ill-defined tasks.

Design Datasets with Ambiguous or Implicit Statements. Many current datasets are too easily solved with keyword-based heuristics (e.g., detecting terms like “GHG” or “climate change”). However, LLMs struggle with more implicit or ambiguous cases. To assess true model understanding, future datasets should focus on examples that require inference and contextual reasoning.

Include Simple Baselines to Assess Task Difficulty. We observed that many tasks can be solved with simple models such as word frequency or TF-IDF. Including these baselines is essential to gauge the added value of advanced models. Ideally, a subset of the dataset should be challenging for such baselines, indicating that genuine language understanding is required.

Quantify Annotation Error Rates and Their Impact. When model performance exceeds 90%, even a small number of annotation errors can meaningfully distort evaluation. We recommend estimating the annotation error rate through manual review or IAA agreement analysis. This allows researchers to define a performance margin below which observed differences may fall within the noise of the dataset, and thus not reflect genuine model improvement. Such quantification is essential for drawing reliable conclusions from small performance gains. This also highlights the importance of computing meaningful confidence intervals measuring uncertainty of performances due to factors such as dataset sampling or initialization.

Prioritize Dataset Quality Over Quantity. As

model accuracy rises, even small annotation errors can significantly impact evaluation. We recommend focusing on clean, high-quality test sets rather than large weakly labeled datasets, which often lack robust guidelines. Poor LLM performance on such datasets may reflect structural flaws rather than model limitations.

5 Conclusion

In this work, we curated, cleaned, and standardized 29 climate-related NLP datasets, ensuring consistency across tasks. We systematically evaluated a diverse range of models. Our results indicate that all approaches achieved competitive performance. However, since even the TF-IDF baseline performs well, the datasets may overall be too simple. Our analysis also revealed that nearly all datasets contain annotation inconsistencies, which may introduce noise in model evaluations. These findings highlight the importance of dataset quality in benchmarking and call for more rigorous annotation protocols in climate-related NLP research.

All our data, scripts, and models are publicly available at https://github.com/tcalamai/acl_climateNLPtoolbox. In particular, we provide a Python library that can run all models on entire files by parsing the PDFs, splitting them into paragraphs, doing model inference on each paragraph (individually or as an ensemble), and aggregating the results into statistics (example in Appendix D).

Acknowledgement. This work was performed using HPC resources from GENCI-IDRIS (Grant AD011014244R1). This work is supported by Amundi Technology and the ANRT with a CIFRE fellowship.

6 Limitations

Our study has several limitations. First, it relies on publicly available datasets, which may not fully represent the entire scientific corpus of datasets related to climate-related tasks.

Second, while we accounted for variability in training through uncertainty estimation, our results remain influenced by factors such as hyperparameter selection. These factors can significantly impact model performance, and while our findings align with those of previous studies, further optimization could improve performance.

Third, we designed our prompts to maximize model performance under a zero-shot setting with

a small chain-of-thought (CoT) component. However, alternative prompting strategies, such as few-shot learning, self-reflection, and iterative reasoning, could enhance LLM performance. Additionally, our study was limited to smaller language models, and larger models, such as LLaMA 3 405B, could yield improved results.

Regarding annotation, error analysis was conducted by the authors of this study. To facilitate interpretability, we exposed both the gold label and the model’s prediction during the annotation process, allowing us to analyze the model’s reasoning. Despite this, we believe our error analysis provides meaningful insights into model behavior. Moreover, we computed the IAA on a subset of the datasets for 2 annotators; we found a Cohen’s κ 0.395 for the 3 label classification: “actual error”, “debatable”, “misclassified”. This is a weak agreement, making the exact figures mentioned in this paper approximations. However, the agreement is better on classifying errors as “actual errors” or not ($\kappa = 0.591$). Moreover, the disagreement between the annotators was mostly with the “debatable” label. This further confirms that the wrong predictions partly stem from HLV.

Finally, we acknowledge a broader ethical consideration: developing tools that analyze how entities communicate about climate-related issues may inadvertently enable strategic adaptation by these entities. Specifically, organizations could leverage such tools either to evade detection or to optimize their messaging to increase perceived compliance. This risk underscores the importance of continuous evaluation and responsible deployment of NLP systems in this domain.

References

- Alix Auzepy, Elena Tönjes, David Lenz, and Christoph Funk. 2023. [Evaluating tcf reporting—a new application of zero-shot analysis to climate-related financial disclosures](#). *PLOS ONE*, 18(11):1–23.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.
- Julia Bingler, Mathias Kraus, and Markus Leippold. 2021. [Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures](#). *Social Science Research Network*.
- 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 4000708.
- Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769.
- Xavier Bouthillier and Gaël Varoquaux. 2020. *Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020*. Ph.D. thesis, Inria Saclay Ile de France.
- Samuel R. Bowman and George Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.
- Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M. Suchanek. 2025. [Corporate greenwashing detection in text - a survey](#). *Preprint*, arXiv:2502.07541.
- Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021. [Computer-assisted classification of contrarian claims about climate change](#). *Scientific Reports*, 11(1):22320.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims](#). *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020, Online, 11 December 2020 - 11 December 2020*.
- Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.
- David Friederich, Lynn H. Kaack, Alexandra Luccioni, and Bjarne Steffen. 2021. [Automated Identification of Climate Risk Disclosures in Annual Corporate Reports](#). Papers 2108.01415, arXiv.org.
- Eduardo C. Garrido-Merchán, Cristina González-Barthe, and María Coronado Vaca. 2023. [Fine-tuning climatebert transformer with climatext for the disclosure analysis of climate-related financial risks](#). *Preprint*, arXiv:2303.13373.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Odd Erik Gundersen, Kevin Coakley, Christine Kirkpatrick, and Yolanda Gil. 2022. Sources of irreproducibility in machine learning: A review. *arXiv preprint arXiv:2204.07610*.

- Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. [MAFALDA: A benchmark and comprehensive study of fallacy detection and classification](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.
- Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. [Logical fallacy detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023a. [Annotation error detection: Analyzing the past and present for a more coherent future](#). *Computational Linguistics*, 49(1):157–198.
- Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023b. [Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future](#). *Computational Linguistics*, 49(1):157–198.
- John Lang, Camilla Hyslop, Natasha Lutz, Natalie Short, Richard Black, Peter Chalkley, Thomas Hale, Frederic Hans, Nick Hay, Niklas Höhne, Angel Hsu, Takeshi Kuramochi, Silke Mooldijk, and Steve Smith. 2023. [Net zero tracker](#). Energy and Climate Intelligence Unit, Data-Driven EnviroLab, NewClimate Institute, Oxford Net Zero.
- Tong Lin, Tianliang Xu, Amit Zac, and Sabina Tomkins. 2023. [Sustainable signals: An ai approach for inferring consumer product sustainability](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6067–6075. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. [Detecting Stance in Media On Global Warming](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.
- Meta. 2024. [Introducing llama 3.1: Our most capable models to date](#).
- 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*.
- Gaku Morio and Christopher D Manning. 2023. [An NLP benchmark dataset for assessing corporate climate policy engagement](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Gabrijela Perković, Antun Drobnyak, and Ivica Botički. 2024. [Hallucinations in llms: Understanding and addressing challenges](#). In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088.
- Paulo Pirozelli, Marcos M. José, Igor Silveira, Flávio Nakasato, Sarajane M. Peres, Anarosa A. F. Brandão, Anna H. R. Costa, and Fabio G. Cozman. 2023. [Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change](#). *Preprint*, arXiv:2309.10945. <https://github.com/C4AI/Pira>.
- Susanna Rücker and Alan Akbik. 2023. [CleanCoNLL: A nearly noise-free named entity recognition dataset](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8628–8645, Singapore. Association for Computational Linguistics.
- Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. You can teach an old dog new tricks! on training knowledge graph embeddings.
- Rylen Sampson, Aysha Cotterill, and Quoc Tien Au. 2022. [Tcfd-nlp: Assessing alignment of climate disclosures using nlp for the financial markets](#). In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*.
- 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.
- Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. 2023a. [Climatebert-netzero: Detecting and assessing net zero and reduction target](#). *Swiss Finance Institute Research Paper*, (23-110).
- Tobias Schimanski, Chiara Colesanti Senni, Glen Gostlow, Jingwei Ni, Tingyu Yu, and Markus Leippold. 2024a. [Exploring nature: Datasets and models for analyzing nature-related disclosures](#). *SSRN Electronic Journal*.
- Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024b. [ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.
- Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023b. [Bridging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication](#).
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärl, and Denny Zhou. 2023. [Large language models can be easily distracted by irrelevant context](#). *Preprint*, arXiv:2302.00093.

- Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. [Towards answering climate questionnaires from unstructured climate reports](#). *Preprint*, arXiv:2301.04253.
- Dominik Stambach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. [Environmental claim detection](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.
- The World Meteorological Organization (WMO). Wmo confirms 2024 as warmest year on record at about 1.55°C above pre-industrial level. <https://wmo.int/news/media-centre/wmo-confirms-2024-warmest-year-record-about-155degc-above-pre-industrial-level>. [Online; accessed 03-February-2025].
- David Thulke, Yingbo Gao, Petrus Pelsler, 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.
- Saeid Vaghefi, Veruska Muccione, Christian Huggel, Hamed Khashehchi, and Markus Leippold. 2022. [Deep climate change: A dataset and adaptive domain pre-trained language models for climate change related tasks](#). In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*.
- Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. [Towards fine-grained classification of climate change related social media text](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.
- Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. [ClimaText: A Dataset for Climate Change Topic Detection](#). *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020, Online, 11 December 2020 - 11 December 2020*.
- Gengyu Wang, Lawrence Chillrud, and Kathleen McKeown. 2021. [Evidence based automatic fact-checking for climate change misinformation](#). In *ICWSM Workshops*.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. [VariErr NLI: Separating annotation error from human label variation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.
- Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. [ClimateBERT: A Pretrained Language Model for Climate-Related Text](#).
- Vinicius Woloszyn, Joseph Kobti, and Vera Schmitt. 2022. [Towards automatic green claim detection](#). In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '21*, page 28–34, New York, NY, USA. Association for Computing Machinery.
- Kun Xiang and Akihiro Fujii. 2023. [DARE: Distill and Reinforce Ensemble Neural Networks for Climate-Domain Processing](#). *Entropy*, 25(4):643.
- Yinan Yu, Samuel Scheidegger, Jasmine Elliott, and Asa Lofgren. 2024. [climateBUG : A data-driven framework for analyzing bank reporting through a climate lens](#). *Expert Systems with Applications*, 239:122162.

A Datasets

Summary of the construction of the dataset To construct the datasets used for the reproducibility study, we followed the following steps:

1. We collected the datasets: we present statistics on their size in Table 3;
2. We explored each dataset to assess data quality (language, gibberish, text length, language), understand the labels (single label or multi-label, relation classification or multi-class classification);
3. We removed text longer than 4000 tokens, and smaller than 5 tokens, for details on which datasets were affected see Table 4;
4. We cleaned the text for formatting, punctuation, space, and encoding issues, and we removed duplicates: we report the statistics in Table 5 ;
5. We tried to keep existing splits if they are provided in the original source. However, if there was data contamination between the train set and the test set and the instances of contamination are rare, we remove the contamination. If there is a dataset construction issue, we rebuilt the dataset to remove the contamination.
6. We constructed based on the previous steps a training and development dataset, each smaller than 10k examples (as it is intended

for fine-tuning). We also construct a test dataset, however for the test dataset we keep the maximum size (to stay as close as possible to the original paper). During down-sampling, we ensured that label distribution was preserved through stratification. For heavily imbalanced datasets, we adjusted the sampling to improve the balance. In Table 6 we detail the ratio between the most frequent and least frequent label in the datasets, in the original datasets and in the newly created datasets.

Dataset	Size			max. length
	train	dev	test	
ESGBERT action500	400	50	50	139
Green Claims	618	77	78	93
SUSTAINABLESIGNALS reviews	623	78	78	246
Implicit/Explicit Green Claims	618	77	78	93
climateFEVER claim (our split)	1228	154	153	78
logicClimate	679	218	179	316
ESGBERT G	1600	200	200	200
Global-Warming Stance (GWSD)	1890	210	200	70
ESGBERT S	1600	200	200	200
ESGBERT E	1600	200	200	200
esgbert Biodiversity	1760	220	220	445
esgbert Forest	1760	220	220	445
esgbert Nature	1760	220	220	445
esgbert Water	1760	220	220	445
Environmental Claims	2117	265	265	76
climatext (Wiki-doc) 10k	6000	300	300	282
climatext (Wiki-doc) wiki	3000	300	300	282
climate sentiment	800	200	320	698
Climate Specificity	800	200	320	698
Commitments And Actions	800	200	320	698
Net-Zero/Reduction	2752	345	344	1436
climateEng	2871	354	355	345
climateStance	2871	354	355	345
ClimateBERT’s Climate detection	1040	260	400	701
Climate TCFD recommendations	1040	260	400	701
climateFEVER evidence	6140	765	770	441
climatext (Wiki-doc) claim	6000	300	1000	282
sciDCC	9224	1154	1161	503
ClimaINS	13755	1710	1710	5426
CC-Contrarian Claims	23436	2605	2898	4677
LobbyMap (Query)	11728	1320	3817	16743
climatext (Wiki-doc)	115847	3618	3826	317
LobbyMap (Stance)	15038	1660	4718	16743
climaQA	71478	8934	8936	1065
ClimaTOPIC	46803	8771	8984	1056
LobbyMap (Pages)	67091	7289	15755	16743
climateBUG-data	96852	24214	29551	4567

Table 3: Original datasets: length of each split (train, test, dev) for each dataset, and size of the largest examples for each dataset (in number of tokens).

Duplicates in ClimaQA ClimaQA’s original construction method inadvertently produced many duplicates. The task of ClimaQA is to determine whether a given a CDP response answers a specific CDP question. The dataset contains correct response-question pairs (positive pairs) and mismatching response-question pairs (negative pairs). For the negative pairs, responses were randomly selected from the pool of all other responses, exclud-

Dataset	>4000 tokens	
Lobbymap (Pages)	66	0.073%
Lobbymap (Query)	48	0.285%
Lobbymap (Stance)	65	0.304%
ClimaINS	18	0.105%
CC-Contrarian Claims	1	0.003%
ClimateBUG-data	2	0.001%

Table 4: Number of text larger than 4000 tokens

ing the correct one. This naive approach led to mis-labeled examples due to all companies responding to the same set of questions. Additionally, the issue of duplication is exacerbated as companies often reuse or copy responses across multiple questions, increasing the likelihood of selecting an appropriate answer as a negative example. The original ClimaQA dataset led to poor performances; we therefore rebuilt the dataset making sure to avoid those issues.

Duplicates in ClimateFEVER For ClimateFEVER, the original dataset does not have a split. However, in ClimaBench (Spokoyne et al., 2023), the authors created the split at the evidence level. The same claim can be found in the training and the testing split but associated to different evidences. This resulted in partial contamination. We experimented with both evidence and claim -based splits (ClimateFEVER (climabench split) and ClimateFEVER (our spl)).

B Experimental Settings

Each dataset was divided into train, test, and development splits. When already available, we kept the initial split; otherwise, we used proportions of 80%, 10%, 10%. The dataset construction is detailed in Appendix A. The datasets with original test split are **climatext (wiki-doc, wikipedia, 10k, claim)**, **Environmental Claims**, **CC-Contrarian Claims**, **Lobbymap (Pages, Query, Stance)**, **ClimateEng**, **ClimateStance**, **LogicClimate**, **ClimaTOPIC**. For ClimaINS and ClimaQA we experimented with both the original splits and our splits.

To establish a benchmark for comparison, we computed several baselines using different statistical and machine-learning approaches:

Random Baseline A model that randomly predicts the label (following a uniform distri-

dataset	text duplicates	cleaned text duplicates	exact duplicates	Dataset Size
ESGBERT E	61	79	73	2000
CC-Contrarian Claims	20	32	30	28939
climateBUG-data	17445	20965	18815	150617
ClimaINS	2538	6986	6778	17175
climateFEVER evidence	3473	3473	2625	7675
logicClimate	458	458	113	1351
Net-Zero/Reduction	0	14	14	3441
Green Claims	0	0	0	773
LobbyMap (Stance)	7837	7843	4365	21416
Commitments And Actions	0	0	0	1320
climaQA	74190	74584	33580	89348
Environmental Claims	0	0	0	2647
SUSTAINABLESIGNALS reviews	4	4	2	779
esgbert Biodiversity	79	79	69	2200
climateStance	0	6	4	3580
Implicit/Explicit Green Claims	0	0	0	773
esgbert Water	79	79	79	2200
esgbert Nature	79	79	79	2200
climatext (Wiki-doc)	2597	2650	2595	123291
climatext (Wikipedia)	49	55	55	3600
climatext (10k)	299	327	327	6600
climatext (claim)	292	322	322	7300
esgbert Forest	79	79	73	2200
climateEng	0	6	4	3580
sciDCC	87	87	37	11539
LobbyMap (Pages)	2182	2303	2067	90135
Climate TCFD recommendations	0	0	0	1700
ESGBERT G	36	54	45	2000
ClimaTOPIC	42	717	675	64558
ESGBERT action500	19	21	21	500
Global-Warming Stance (GWSD)	266	270	268	2300
ESGBERT S	5	23	23	2000
ClimateBERT’s Climate detection	0	0	0	1700
climateFEVER claim (our split)	0	2	2	1535
climate sentiment	0	0	0	1320
Climate Specificity	0	0	0	1320
climateFEVER evidence	3473	3473	2625	7675

Table 5: Number of duplicates in the unfiltered datasets. *text duplicates* correspond to duplicates of the unprocessed texts, *cleaned text duplicates* correspond to duplicates of text after the formatting and encoding cleaning, *exact duplicates* correspond to duplicates of both the input text and the label. Therefore, the difference between *cleaned text duplicates* and *exact duplicates* correspond to duplicates with mismatching labels.

dataset	Imbalance ratio			Weighted Training
	Original	Filtered	Test	
climaQA	1.0	1.0	1.0	
climatext (Wiki-doc)	1.0	1.0	1.0	
esgbert Nature	1.1	1.1	1.0	
SUSTAINABLESIGNALS reviews	1.3	1.3	1.3	
ESGBERT action500	1.4	1.4	1.3	
Commitments And Actions	1.4	1.4	2.3	
ESGBERT S	1.5	1.5	1.5	
Climate Specificity	1.5	1.5	1.6	
Net-Zero/Reduction	1.6	1.6	1.6	
climate sentiment	1.6	1.6	3.2	
ClimaINS	1.7	1.7	2.1	
Green Claims	1.9	1.9	1.9	
ESGBERT E	2.0	2.1	2.2	
Global-Warming Stance (GWSD)	2.2	2.2	2.3	
climateBUG-data	2.3	2.3	1.3	
LobbyMap (Stance)	2.5	2.3	2.7	
ESGBERT G	2.7	2.7	2.8	
Environmental Claims	3.0	3.0	3.0	
esgbert Water	3.1	3.1	3.4	
ClimateBERT’s Climate detection	3.3	3.3	4.0	
esgbert Forest	4.0	4.1	4.3	
climateFEVER claim (our split)	4.3	4.2	4.0	
Implicit/Explicit Green Claims	4.3	4.3	4.2	
esgbert Biodiversity	4.7	4.8	4.9	
LobbyMap (Pages)	4.7	4.5	3.0	
climateFEVER evidence	6.3	2.4	5.6	✓
Climate TCFD recommendations	8.3	8.3	7.6	✓
climateStance	8.5	8.5	8.1	✓
climatext (wiki)	10.5	10.3	8.1	
climateEng	12.0	12.1	23.1	✓
climatext (claim)	17.8	17.2	1.0	
climatext (10k)	17.8	17.2	3.5	
logicClimate	34.8	34.8	45.0	✓
ClimaTOPIC	44.3	3.4	26.7	
sciDCC	49.2	48.8	49.5	✓
LobbyMap (Query)	65.3	65.3	47.0	
CC-Contrarian Claims	110.9	19.5	83.2	✓

Table 6: Ratio between the most frequent and least frequent label in the dataset. The ratio is reported for the original dataset, then after filtering for token length and dataset size, as well as for the test set. For heavily imbalanced datasets, we used a weighted loss during training.

tribution) implemented with sklearn’s DummyClassifier.

TF-IDF A logistic regression using TF-IDF features to predict the class, implemented with sklearn’s LogisticRegression and TfidfVectorizer.

DistilRoBERTa A DistilRoBERTa (Sanh et al., 2019) fine-tuned on each task implemented with transformers’ AutoModelForSequenceClassification for 10 epochs, with AdamW optimizer and a learning rate of 5e-5.

Longformer A Longformer (Beltagy et al., 2020) fine-tuned on each task implemented with transformers’ AutoModelForSequenceClassification for 10 epochs, with AdamW optimizer and a learning rate of 5e-5.

GPT-4o-mini GPT-4o-mini used as a zero-shot

classifier implemented with OpenAI’s API (temperature of 0.1, top_p of 1)

Llama 3.1 8B Llama 3.1 8B used as a zero-shot classifier (temperature of 0.1, top_p of 1)

Llama 3.1 70B Llama 3.1 70B used as a zero-shot classifier (temperature of 0.1, top_p of 1)

Additionally, we reported the performances reached in the original studies. Compiling and comparing reported performances from various papers revealed significant heterogeneity in performance metrics (Calamai et al., 2025). We reported the value in the result tables; however, they are not necessarily comparable.

Finetuning Parameters We finetuned for each dataset a Longformer and a distilRoBERTa for 10 epochs (with early stopping based on the validation F1-score). To speed up the training, we used half-precision floating-point (fp16). To maximize the GPU resources, we used a batch of 16 for distilRoBERTa and 7 for Longformer (with an accumulation step of 2). We set the seed/random state to 42 for every library (both to create the dataset and train the models). We tried setting parameters to deterministic, however, Longformer was not compatible with PyTorch deterministic algorithms. The models are fine-tuned on the training dataset, evaluated at each epoch on the development dataset, and the best model is selected using the development dataset. We used a learning rate of 5e-5 (warmup ratio of 0.1, a weight decay of 0.01) following Spokorny et al. (2023)’s parameters choices.

Baselines For the simple baselines, we used the seed 42. We used the scikit-learn implementations of the random classifier, the TF-IDF vectorizer, and the logistic regression. For the logistic regression, we used the weights to handle imbalanced datasets. For the relation classification, we parse both texts using the TF-IDF vectorizer and then concatenate the 2 vectors for the logistic regression. The model is fitted on the training dataset.

Zero-shot For the zero-shot models, we collected the description of the labels from each paper and a description of the annotation task. When missing, we wrote a description. We then designed a corresponding prompt template. Using GPT-4, we generated all the instruction prompts for each dataset. We then generated the prediction for a small subset of each dataset (50 examples). Using the outputs,

we validated the format of the prompts and ran the experiment on a larger dataset. We identified issues with some prompts that were manually updated. Finally, we experimented with both zero-shot and Chain-of-thought (CoT). If the performances on the subset were significantly improved, we selected the CoT prompt. The results using the CoT prompts are: *climateFEVER*, *climateStance*, *climatext (10K, wiki, claims, wiki-doc)*, *GWSD*, *LobbyMap*.

C Detailed Analysis of Performance

In this section, we deep dive into the performances of models and errors analysis for each dataset. For each task, we provide a general result analysis, both on the performance of the models and on the error annotation that we conducted.

C.1 Climate-Related Topic Detection

Task Description Given an input sentence or a paragraph, output a binary label, “*climate-related*” or “*not climate-related*”.

Experiment We collected all available datasets: *climateBUG-Data* (Yu et al., 2024), ClimateBERT’s *climate detection* (Bingler et al., 2023), *ClimaText* (Varini et al., 2020), and *SUSTAINABLESIGNALS reviews* (Lin et al., 2023). *Climatext* is composed of *climatext wiki-doc* composed of text weakly labeled; and *climatext wikipedia*, *10k* and *claim* which are annotated datasets. For our experiments, we trained our models on *climatext wikipedia* training split as it was the larger annotated one. For *climatext 10k* we also used the training split for 10-Ks which contains only 58 positive samples. *climatext claim* did not contain a training split, so we used the same as for *climatext 10k*. Our cleaning and processing steps affected the *climateBUG-data* and all *climatext* datasets by removing duplicates, smaller/larger texts, and encoding issues. The results of the experiments are shown in Table 8.

Analysis We found that the TF-IDF baseline achieves strong performance, with values between 65.8% and 86.4%. This is expected as some words are heavily associated with climate change such as “greenhouse gas” or “climate”. However, the baseline is outperformed by the fine-tuned transformers, with macro F1-scores between 73.7% and 95.8%. This difference is statistically significant for all datasets except *SUSTAINABLESIGNALS reviews*. **This shows that (1) the task relies on vocabu-**

dataset	Errors in sample			Overlapping Errors				
	8B	70B	GPT	8B ∩ GPT	8B	70B ∩ GPT	70B	GPT
climate sentiment	9	7	10	56	84	54	86	68
Environmental Claims	12	11	14	26	43	26	57	29
Climate TCFD recommendations	7	7	10	179	230	166	207	211
Implicit/Explicit Green Claims	5	9	12	5	16	9	16	12
climatext (10k)	8	5	9	8	17	5	8	9
Climate Specificity	8	11	16	58	75	75	102	91
Green Claims	4	2	6	4	7	2	10	6
climateBUG-data	17	14	20	81	238	74	128	112
ClimalNS (our split)	9	9	11	123	195	117	172	142
Global-Warming Stance (GWSD)	10	7	10	48	71	35	56	61
ESGBERT action500	8	7	10	8	15	7	17	10
sciDCC	9	10	10	597	750	573	670	646
ClimateBERT’s Climate detection	11	12	17	11	33	12	16	17
climatext (Wikipedia)	14	10	14	16	18	12	14	16
climatext (Wiki-doc)	18	15	23	100	123	63	101	145
Net-Zero/Reduction	5	6	8	5	19	6	14	8
climateStance	5	9	10	96	158	88	113	120
CC-Contrarian Claims	7	8	10	359	588	322	473	424
Commitments And Actions	15	10	20	84	110	82	157	100
esgbert Forest	9	7	10	9	16	7	19	10
esgbert Water	9	4	9	9	15	4	10	9
climateFEVER evidence	7	8	10	218	357	182	251	286
ClimaTOPIC	10	10	10	424	496	424	461	471
ESGBERT E	5	5	8	5	8	5	9	8
climatEng	8	6	10	61	106	71	99	101
climaQA	13	10	20	96	379	94	222	188
Total	242	219	317	2686	4167	2515	3488	3300

Table 7: This table presents the overlap between errors made by GPT-4o-mini and Llama models (8B and 70B). The first two columns shows the number of errors of Llama models among the sample of error of GPT-4o-mini (third column). The overall overlapping errors between GPT-4o-mini and Llama (8B) / Llama (70B) are reported in the following columns, along with the total number of errors for each model. (We do not include *LogicClimate* and *LobbyMap* because they are multilabel datasets)

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
climateBUG-data(Yu et al., 2024)	49.6(49.0-50.1)	86.4(86.0-86.8)	90.5(90.1-90.8)	90.5(90.2-90.8)	89.2(88.8-89.5)	79.0(78.5-79.4)	88.3(87.9-88.6)	91.36 ^f
ClimateBERT Cl. det.(Bingler et al., 2023)	42.8(38.0-47.8)	79.3(74.3-83.8)	95.8(93.2-98.0)	94.0(91.0-96.9)	93.2(90.0-96.3)	88.3(84.3-91.8)	93.4(90.1-96.4)	99.13 ^c
climatext (Wiki-doc)(Varini et al., 2020)	49.3(47.8-50.8)	80.4(79.2-81.7)	85.6(84.5-86.8)	83.8(82.6-85.0)	84.0(82.8-85.2)	86.3(85.1-87.3)	89.8(88.8-90.7)	-
climatext (10k)(Varini et al., 2020)	46.9(41.5-52.2)	91.0(86.9-95.0)	97.0(94.3-99.0)	96.5(93.6-98.7)	95.4(91.8-98.2)	90.9(86.3-94.8)	96.0(93.0-98.5)	95.5 ^{1,f}
climatext (claim)(Varini et al., 2020)	48.2(45.1-51.2)	74.7(71.7-77.3)	67.8(64.6-70.9)	75.4(72.5-78.2)	82.2(79.7-84.6)	83.6(81.4-85.8)	85.7(83.7-87.8)	83.5 ^{1,f}
climatext (Wikipedia)(Varini et al., 2020)	40.1(35.3-44.9)	83.5(75.0-90.4)	83.9(76.0-90.4)	88.2(81.0-94.0)	84.7(77.1-91.5)	82.8(74.8-89.9)	87.4(80.4-93.5)	80.2 ^{1,f}
Sustainable Signals rev.(Lin et al., 2023)	37.6(27.0-49.0)	65.8(54.4-76.6)	73.2(63.2-83.3)	73.7(62.9-82.7)	69.7(57.6-79.7)	63.5(52.5-73.9)	73.3(62.6-82.4)	-

Table 8: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *1.* binary F1-score, *3.* weighted average F1-score, *c.* climateBERT, *f.* climateBUG-LM, *t.* BERT.

lary, but (2) vocabulary is not enough to identify perfectly if the statement is about climate.

All zero-shot approaches reach performances above 75% showing that they can identify text as climate-related. The best zero-shot approaches even reached performances similar to the fine-tuned models. GPT-4o-mini and Llama 3.1 (70B) outperformed the fine-tuned models on *climaText (Wiki)*, and slightly underperformed on *climateBUG-data*, *ClimateBERT’s Climate Detection* and *SUSTAINABLE SIGNALS reviews*. However, the difference between fine-tuning and zero-shot is not significant, except for Llama 3.1 70B outperforming significantly all fine-tuned models on *climaText (Wiki)*.

The reference performances remain higher, with F1-scores above 90% (Garrido-Merchán et al., 2023; Bingler et al., 2023; Yu et al., 2024).

C.1.1 Error analysis

climatext (Wiki-Doc) When classifying sentences as climate-related or not, 91% of errors from GPT-4o-mini are FN. There are only 13 FP. 7 of them come from pages on topics that are indeed somewhat related to climate change, but that do not count as climate-related in Climatext’s methodology – for example, pages about lakes that will be impacted by climate change (Example C.1). Of the remaining FP, 4 are out-of-context statements and 2 are actual errors from GPT-4o-mini. As for the sampled FN, 6/10 are text that is only indirectly linked to climate change. For example, a page about the lithosphere (Example C.2) counts as climate-related in climatext’s methodology, but not in GPT-4o-mini’s view. In 4/10 FN are out-of-context statements. For example, a statistics about politics (Example C.3) counts as “climate-related”

in *climatext*'s methodology, but not in the view of GPT. **It is important to note that the *climatext* (Wiki-Doc) is weakly labeled, therefore the actual task is to identify if a sentence is extracted from a Wikipedia page related to climate change.** Some sentences are therefore out-of-context and impossible to correctly classify. This dataset could be better used as a only a training dataset.

Example C.1 (Not Climate-related). IISD Experimental Lakes Area (IISD-ELA) is a natural laboratory consisting of 58 small lakes and their watersheds set aside for scientific research.

Example C.2 (Climate-related). There are two types of lithosphere:

Example C.3 (Climate-related). The Senate vote throughout the states was between 10 and 20 percent.

ClimaText (Wikipedia/10-Ks) For *climatext* (Wikipedia) and *climatext* (10-Ks) the majority of errors are FN (respectively 75% and 89%).

There are actually only 4 FP for *climatext* (Wikipedia) and 1 for *climatext* (10-Ks). All the FP are linked to environmental issues, and depending on the context, could be linked to climate change. Either directly such as Example C.5 or indirectly such as Example C.4. We believe those errors are all ambiguous and debatable.

For the FN in *climatext* (wiki), we found 3 examples that are annotation errors (e.g. Example C.6 as “Acronyms, potentially well connected to climate change, must be mentioned along with some mechanism/cause/effect of climate change”). Those annotation errors occurred on examples that are ambiguous, yet precisely described in the guidelines. *This shows that Human do also struggle with ambiguous statements.* We found 2 errors that are due to our prompt which is a simplified version of the guidelines, therefore missing some nuance. 3 are due to indirect links to climate-change. The remaining 2 sampled errors are debatable. In the guidelines, it is mentioned that “*Just mentioning clean energy, emissions, fossil fuels, etc. is not sufficient: rather it must be connected to an environmental (CO2) or societal aspect (divestment, Kyoto treaty) of climate change.*”. Example C.7 was classified by the model as “not climate-related” while the annotator decided it was “climate-related”. While this example is definitely climate-related with the use of “zero-carbon”, when strictly following the guidelines, this could be labeled “not climate-related”.

For the FN in *climatext* (10-Ks), due to the high performances, there are actually only 8 FN. There are 2 actual errors due to the following reasoning: “*However, it does not directly discuss the mechanisms, causes, or effects of climate change itself*” (Example C.8). *The model was anchored to the part of the prompt mentioning mechanisms, causes, or effects.* 2 errors are debatable due to the issue mentioned for *climatext* (wiki). The remaining 4 errors are obvious annotation errors, most likely due to the Active Learning approach used to create the dataset. Those examples contain words such as “political climate” or “economic climate” (e.g. Example C.9). They were mislabeled in the dataset but correctly classified by GPT-4o-mini.

Example C.4 (Not Climate-related). Think-tanks such as the World Pensions Council (WPC) argued that the keys to success lay in convincing officials in the U.S. and China, by far the two largest national emitters:

Example C.5 (Not Climate-related). “Indirect effects include the fact that aerosols can act as cloud condensation nuclei, stimulating cloud formation.”

Example C.6 (Climate-related). The location of UNFCCC talks is rotated by regions throughout United Nations countries.

Example C.7 (Climate-related). As of 2012, France generated over 90% of its electricity from zero carbon sources, including nuclear, hydroelectric, and wind.

Example C.8 (Climate-related). In addition, we cannot control the actions of our third party manufacturers or the public’s perceptions of them, nor can we assure that these manufacturers will conduct their businesses using climate change proactive or sustainable practices.

Example C.9 (Climate-related). The current economic climate, especially in Europe, may have an adverse effect in the markets in which we operate.

ClimateBERT’s climate detection For ClimateBERT’s climate detection dataset, there are 11 FP and 6 FN. The majority of FP (7/11) stem from the use of the ambiguous word “sustainable”. Indeed, “sustainable” can be used in a business context meaning that the business is economically viable (Example C.10). However, some of those FP (3) mentioning “sustainable” are actually debatable (e.g. Example C.11 in which “sustainable concepts” is likely to refer to environmental sustainability).

In the remaining FP, 2 are actually mislabeled, and 1 is debatable.

Of the 6 FN, 4 are indirectly but not explicitly linked to climate change or the environment. For example, the dataset labels as climate-related some statements about the fossil fuel/energy industry that do not mention anything regarding the environmental impact, policy, or governance (Example C.12). It also labels as positive statements about the governance of sustainability-related activity (Example C.13).

Example C.10 (Not Climate-related). With passion and integrity at the heart of everything we do, we aim to build a sustainable business that makes a positive difference for all, now and in the future.

Example C.11 (Not Climate-related). We have a long heritage of innovation and strive to provide athletes with the best by creating high-performance and competitive products. In 2020, we continued to serve consumers with innovative technologies and sustainable concepts built into our products.

Example C.12 (Climate-related). On January 5, 2021, Centrica plc. closed a transaction to sell its entire ownership interest in Direct Energy to NRG Energy Inc. (NRG). Effective January 5, 2021, NRG provided a \$300 million guarantee, supported by a \$300 million letter of credit for Direct Energy’s obligations to ATCO Gas and ATCO Electric under the transaction agreements.

Example C.13 (Climate-related). The diagnosis of organizational culture represents the internal scenario and it is one of the elements considered/analyzed in the definition of the drivers. Since 2017, after reviewing the research strategy by our organizational behavior area, only the Organizational Climate and Engagement Survey was applied annually, as presented in the table below.

ClimateBUG In this dataset, GPT produces FN (54%) and FP (46%) in roughly equal proportion. In the sampled errors, we found that 2 FN are statements where the link to climate comes from the context – which is however not given (Example C.14). As in *ClimateBERT’s climate detection*, the word “sustainability” caused problems. However, in this case, the annotation guidelines are still ambiguous. It mentions that “Sustainability not related to the environment (i.e. sustainable profits)” should be classified as “Not Climate”. However, when not given sufficient context, the word can still be ambiguous. Some examples mention “sustainability”

in a broader sense (without explicitly mentioning environmental/climate sustainability), and they are labeled positive (Example C.15) – although GPT believes they are negative. Vice versa, other broad references to sustainability are labeled as negative (Example C.16) – and GPT believes they are positive. This issue accounts 5/10 FP. This shows that both annotators and GPT struggle with this ambiguity and highlights inconsistencies in the annotations. The remaining FP (5/10) are climate-related tables that GPT-4o-mini classified as positives (therefore not following the annotation guidelines to classify tables as negatives). However, we also found at least 1 table mislabeled in the gold standard (Examples C.17). This also shows the importance of well-defined labels and label consistency during annotation, as it can introduce mislabeled examples. This is particularly important when comparing performances of 88.8% and 90.5%.

Example C.14 (Climate-related). This also gives us the opportunity to exchange knowledge and experiences with representatives from different spheres of society and to implement the principles and reach our targets.

Example C.15 (Climate-related). After all, they are the source of jobs, innovation, sustainability, and prosperity.

Example C.16 (Not Climate-related). Firstly, Nordic customers are relatively advanced in their sustainability considerations, and they expect their bank to be able to support sustainable development across segments.

Example C.17 (Climate-related). SCOPE 2—Electricity consumed (market-based method)—Electricity consumed (location-based method)³, 8 67.

C.2 Climate-Related Subtopic Detection

Task Description Given an input sentence or a paragraph, output a subtopic related to climate change. This is a multiclass classification task.

Experiment We collected the *ClimaTOPIC* (Spokoyny et al., 2023), *esgbert Nature* (Schimanski et al., 2024a), *ClimateEng* (Vaid et al., 2022) and *SciDCC* (Mishra and Mittal, 2021) datasets. The results are in Table 9. Our cleaning and processing steps impacted all datasets: it removed duplicated examples from the *esgbert Nature* and *ClimaTOPIC* test sets, and it removed the URLs in *ClimateEng*. And for *SciDCC*, we made sure that each

category was represented in the test set; however, some categories are severely under-represented (Geology appears only 28 times overall, and therefore only 3 times in the test set). We note that no original split is given for *SciDCC*.

Result Analysis In our experiment, the best-performing models are usually the fine-tuned models (except on Biodiversity label from *esgbert Nature*). While the performance of the fine-tuned models improved over the TF-IDF baseline, the difference is statistically significant only for *ClimaTOPIC*, *esgbert Nature* (nature and water labels). **Overall, this shows that the classes are distinguishable using mostly the vocabulary.**

The zero-shot approaches perform comparably to the fine-tuned transformers on most datasets. While the fine-tuned models outperformed the zero-shot approaches, the difference is statistically significant only on *SciDCC*, *ClimaTOPIC*. Those two datasets are constructed automatically and not annotated by humans; therefore, building a comprehensive prompt is difficult, while identifying linguistic patterns remains effective.

In terms of order of magnitude, our models achieved performance levels comparable to those reported in the original studies. However, we observed a notable performance gap on *ClimaTOPIC* and *SciDCC*, where our models underperformed relative to the results presented in prior work. For *ClimaTOPIC* this could be explained by the smaller training dataset. For *SciDCC*, we provide only the summary of the article, but the dataset also contains the content of the article, which can help increase performance.

C.2.1 Error analysis

SciDCC We have reported the confusion matrix of GPT-4o-mini predictions in figure 1a. We observe several types of errors: (1) hallucinations (“Marine Biology” and “Archaeology” are not labels in this dataset) (2) predicting too frequently the label “Climate” (3) mixing related labels (“biology”, “Biotechnology” and “Genetically Modified”, or “Geology” and “Earthquakes”)

All these labels are so close in meaning because they are categories that changed over time as shown in Figure 1b. We can observe the topic shift over time: “Ozone Holes”, “GMOs” and then “Global Warming”. Categories not only appeared but also changed over time. For example, “endangered animals” topics might have later been included in

“Animals” or “Zoology”. Moreover, multiple categories might fit a specific topic such as “Extinction” and “Climate” as one is the consequence of the other. Moreover, in the original paper, they did not provide a clear definition of each label. Examples of surprising gold labels include: Example C.18 classified as “Geography” and not “Pollution”, or Example C.19 classified as “Geography” and not “Geology”.

When reviewing sampled errors by GPT-4o-mini, we identified that 11/20 are due to text with multiple possible labels (Examples C.21 which could fit in “Animals”, “Endangered Animals”, “Extinction” or “Climate”), 7/20 contain debatable labels (Examples C.20 labeled “Earthquakes” but could fit in “Geology”). The latter can be explained as in our experiment we only provide the summary, but the article’s content might be linked to the original label. **While this dataset might be interesting for studying the evolution of environmental topics, it would require human annotations to re-label the data to make it reliable.**

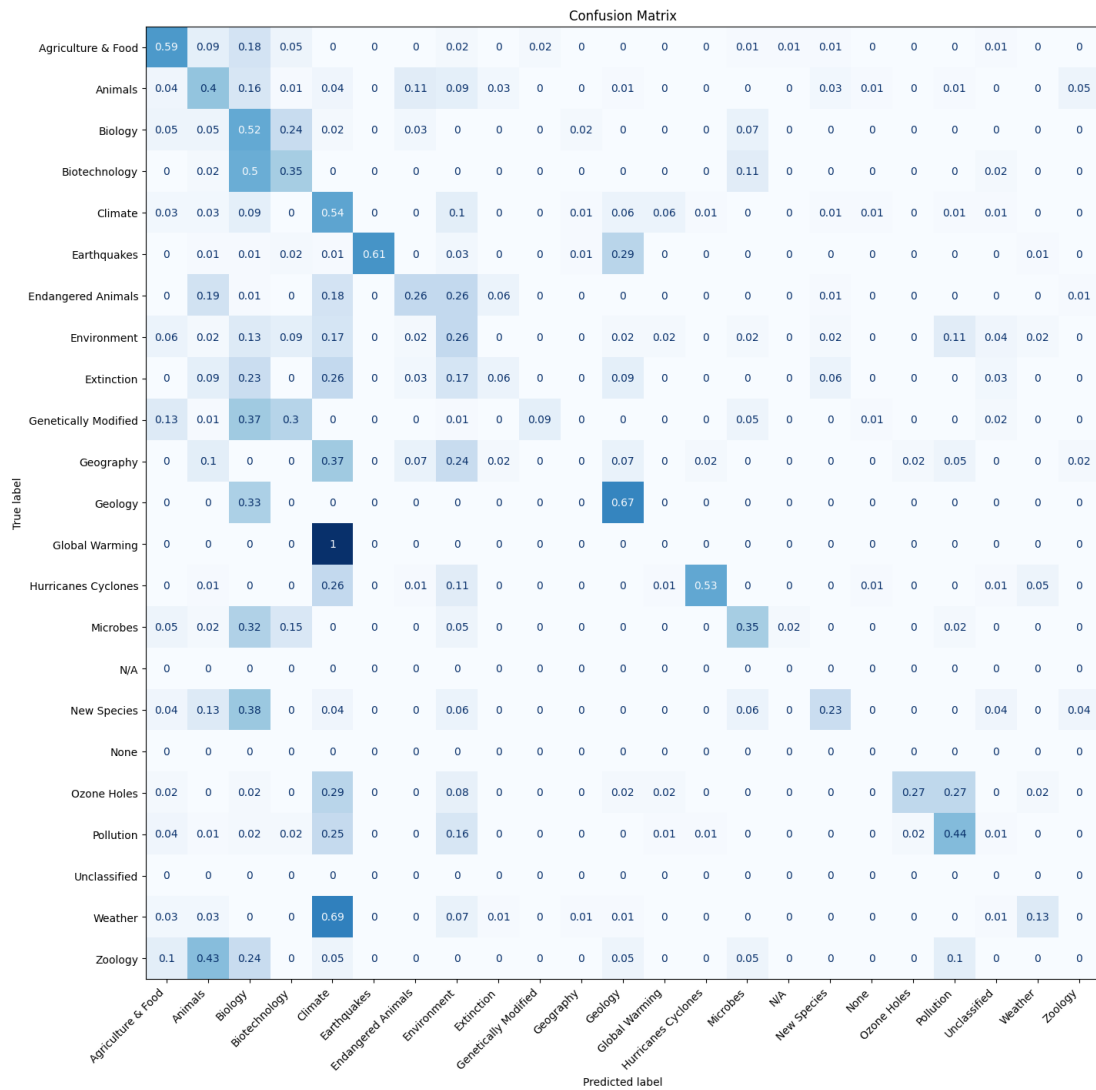
Example C.18 (Geography). Shipping traffic can be a major source of tiny plastic particles floating in the sea, especially out in the open ocean. In a paper published in the scientific journal

Example C.19 (Geography). New research led by the University of Cambridge has found rare evidence – preserved in the chemistry of ancient rocks from Greenland – which tells of a time when Earth was almost entirely molten.

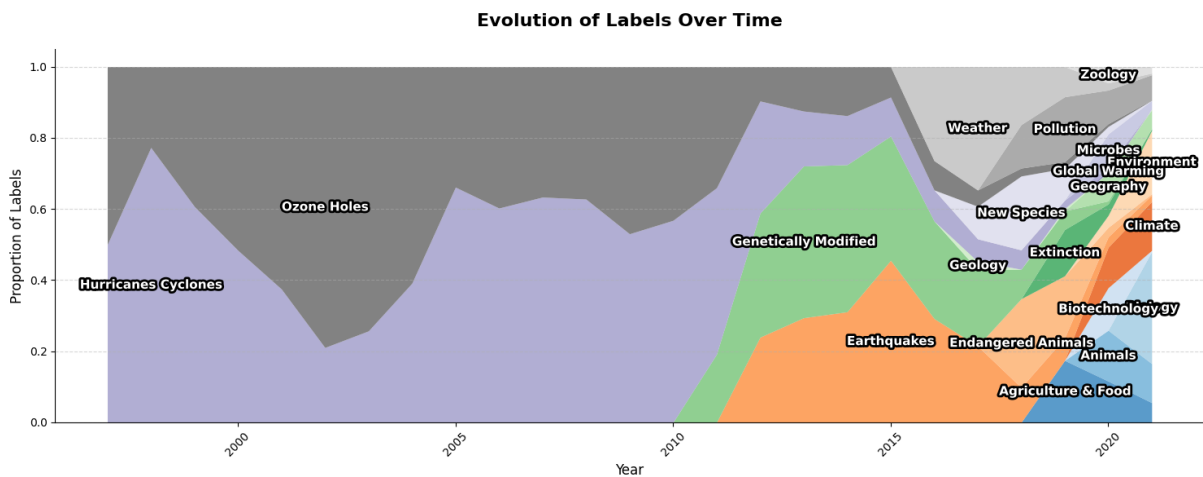
Example C.20 (Earthquakes). Diamonds, those precious, sparkling jewels, are known as the hardest materials on Earth. They are a high-pressure form of carbon and found deep in the ground.

Example C.21 (Animals). Many species might be left vulnerable in the face of climate change, unable to adapt their physiologies to respond to rapid global warming. According to a team of international researchers, species evolve heat tolerance more slowly than cold tolerance, and the level of heat they can adapt to has limits.

ESGBERT Nature GPT-4o-mini performs well overall, but when errors occur, they tend to be FN (90% for forest, 77% for water and 60% for biodiversity), especially in cases involving indirect references. There are 1 FP and 9 FN for the “forest” dataset. All 9 FN were indeed related to forests, but they were misclassified due to a lack of explicit references to forest management. Instead,



(a) Confusion matrix of GPT-4o-mini predictions on *SciDCC* dataset. “Archeology” and “Marine Biology” are not labels from *SciDCC*.



(b) Evolution of proportion of labels over time in *SciDCC* dataset (Mishra and Mittal, 2021)

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
ClimaTOPIC(Spokoyko et al., 2023)	6.0(5.6-6.5)	46.6(45.4-47.8)	55.8(54.4-57.2)	54.8(53.5-56.1)	35.8(34.8-36.7)	34.5(33.4-35.4)	30.5(29.5-31.5)	65.22 ^r
climateEng(Vaid et al., 2022)	13.2(10.4-16.2)	58.4(48.2-66.8)	70.5(60.8-77.7)	67.3(58.3-75.4)	65.8(55.5-73.9)	53.9(43.5-62.2)	61.1(51.9-68.9)	74.58 ^r
esgbert Biodiversity(Schimanski et al., 2024a)	41.1(35.0-47.7)	91.3(85.3-96.2)	91.3(85.6-95.7)	89.0(83.3-94.2)	91.8(86.4-96.2)	88.8(83.0-94.3)	82.3(75.9-87.7)	92.29 ^{l-g}
esgbert Forest(Schimanski et al., 2024a)	41.4(35.4-47.2)	92.9(87.4-97.1)	97.2(93.8-99.3)	97.0(93.8-99.4)	91.6(86.4-96.2)	87.4(81.2-92.5)	86.2(80.3-91.8)	95.37 ^{l-g}
esgbert Nature(Schimanski et al., 2024a)	51.5(45.0-58.0)	82.9(77.7-87.7)	89.8(85.8-93.6)	89.9(85.8-93.8)	82.9(77.6-87.9)	86.7(82.0-91.0)	88.0(83.3-92.4)	94.19 ^{l-g}
esgbert Water(Schimanski et al., 2024a)	42.7(36.6-49.5)	84.6(77.5-90.8)	95.3(91.7-98.6)	93.9(89.7-97.5)	93.8(89.4-97.6)	89.4(83.4-94.1)	93.5(89.1-96.9)	95.10 ^{l-c}
sciDCC(Mishra and Mittal, 2021)	3.7(2.7-4.8)	42.6(39.0-46.7)	39.8(37.1-42.7)	42.1(39.2-45.9)	29.2(26.8-31.4)	20.5(18.3-22.7)	30.0(27.5-32.2)	54.79 ^l

Table 9: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *l*. binary F1-score, *c*. climateBERT, *g*. EnvironmentalBERT, *l*. Longformer, *r*. RoBERTa

these statements included discussions about plants, wood usage, the economic impact of wildfires, and the effects of climate change—topics inherently connected to forestry (Example C.22).

Example C.22. While gross loss reserve estimates for the 2018 California wildfires were also reduced, this was largely offset by a reduction in reinsurance recoverables resulting in very little change to estimated net losses from those wildfires. [Forest]

There are 2 FP and 7 FN for the “water” dataset. The 2 FP are both debatable as they explicitly talk about water-related topics (Example C.23). As part of the FN, we found 2 focusing on financial consequences, therefore indirectly linked to “water”. The “water” label: “The topic of water centers around water management, consumption, and pollution. [...]” could be understood as focusing only on freshwater, and not include water-related natural disasters such as tsunami. We found that 4 FN are related to those topics (Example C.24). The last FN is an actual error by GPT-4o-mini.

Example C.23. 70% of the water in the world is used for agriculture. [Not Water]

Example C.24. In addition, Intel employees responded to these events with great generosity, contributing \$1.8 million toward tsunami relief, \$1.6 million to help the victims of Hurricane Katrina and \$475,000 for those affected by the earthquake in Pakistan.

While this dataset is well-built, these errors highlight the need to have precise definitions and the ambiguity of indirect relation to the topic. This is further emphasized by the inter-annotator agreement (Fleiss’ $\kappa = 80\%$) showing some disagreement between annotators.

ClimateEng Firstly we see that the performance of the GPT-4o-mini model is close to the performances of the finetuned models. We have reported the confusion matrix in figure 2, we see that a large

portion of issues come from the labels “General” and “Politics” being mixed-up.

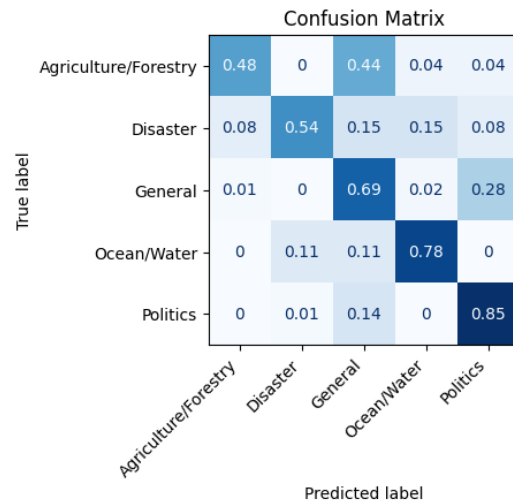


Figure 2: Confusion matrix of GPT-4o-mini predictions on *ClimateEng* dataset.

When investigating the sampled errors, we found that 12/30 are actual errors, 4/30 are debatable and 12/30 are mislabeled examples (Example C.25). We found that the actual errors are related to the very specific definitions of the labels. For example, the political label focuses only on leaders, political organizations and awareness about climate change and not on political activism or opinions. However, we see inconsistencies in the annotations for political awareness: Example C.27 was annotated as “Politics” - while GPT-4o-mini predicted “General” - and the Example C.28 was annotated as “General” - while GPT-4o-mini predicted “Politics”. 2 tweets contain hashtags about COP25, which could fit into the “Politics” category but were annotated as “General”. In the guidelines, the “Ocean/Water” label is specific toward “biodiversity”, but GPT-4o-mini predicted Example C.26 as “Ocean/Water”, therefore not following the instructions. Overall, we observe that the labels are really specific and

counter-intuitive compared to the usual definition of the categories. *GPT-4o-mini* tend to struggle with this re-definition of terms. However, the inter-annotator agreement (Cohen’s $\kappa = 0.739$) also shows that people do have disagreements and also struggle with those definitions. Once again, this shows the importance of having well-defined labels.

Example C.25 (Agriculture/Forestry). Contracts awarded for Scottish offshore wind farm <URL> #energy #sustainability #climatechange

Example C.26 (General). Scientists fear that a change in ocean circulation could profoundly alter climate patterns. <URL>

Example C.27 (Politics). U.S. students excel in global warming awareness, social justice activism, racial angst, and gender sensitivity while China’s teens waste their time gaining proficiency in reading, math and science, according to global education study. <URL>

Example C.28 (General). On December 11, 2019, the Intergenerational Dialogue was successfully held with the participation of Youth Delegation and a series of climate experts. The aim of the dialogue is to exchange personal experience and guide the delegates to the most fertile areas of climate change. <URL>

ClimaTOPIC As the topics are categories of questions from CDP questionnaires, many categories are actually related. The Example C.29 and C.30 both deal with emissions of buildings, however, one is from the “Emissions” category and the other is from the “Building” category. This might be improved by giving more context to the model, in particular, the type of questions available in each category.

From the sample of errors, we found that 9/10 errors are similar mistakes. From the confusion matrix (Figure 3), we see that the model fails most often on “Governance and Data Management”, “Opportunities” and “Strategy” labels which are non-specific categories. While we cannot consider the data to be mislabeled—since the labels correspond to the section headers from which the responses were extracted—it is evident that the model’s predictive performance is constrained by the broad and non-specific nature of these labels.

Example C.29 (Emissions). The Direct emission from institutional buildings occur but we are not

able to estimate it on this inventory, due to financial constraints and limited manpower.

Example C.30 (Building). So as to decrease GHG emissions, under the climate action plan for reducing GHG emissions, the target of the energy upgrade of municipal buildings has been set, tackling energy efficiency. The emission reduction target has been set to 27,804 tnCO₂eq annually.

C.3 Detecting Climate-Related Financial Disclosure

Task Description This is a particular case of sub-topic classification. Given a paragraph or a sentence, output the TCFD recommendation associated with the content of the text.

Experiment We were not able to collect any of the datasets; however, the authors of [Bingler et al. \(2021\)](#) released a smaller TCFD annotated dataset alongside their follow-up work ([Bingler et al., 2023](#)). We identify it as *Climate TCFD recommendations*. So we evaluated this dataset with the setting describe in section B and reported the results in Table 10. Our pre-processing did not alter this dataset.

Result Analysis The best performing approaches were the fine-tuned transformers reaching F1-score above 68%, significantly out-performing the zero-shot approaches and the TF-IDF baseline. The TF-IDF baseline is performing relatively well given the nature of the task. Moreover, it performed better than the zero-shot approaches.

As this dataset is not part of a previous study, we rely on other works for performance comparison. [Bingler et al. \(2021\)](#) observed that paragraph-level approaches underperformed compared to methods that aggregate sentence-level features, such as majority voting or logistic regression. In contrast, [Sampson et al. \(2022\)](#) found that classical models, such as TF-IDF-based random forests and stacked approaches, outperformed fine-tuned transformers, which performed well but not optimally. Interestingly, the zero-shot approach evaluated by [Auzepy et al. \(2023\)](#) yielded surprisingly strong results given the complexity of the task.

C.3.1 Error analysis

climate TCFD recommendations The results, presented in the confusion matrix in Figure 4, revealed that 62% of the model’s classification errors are associated with the “Strategy” label. Additionally, the model tends to over-predict the “None”

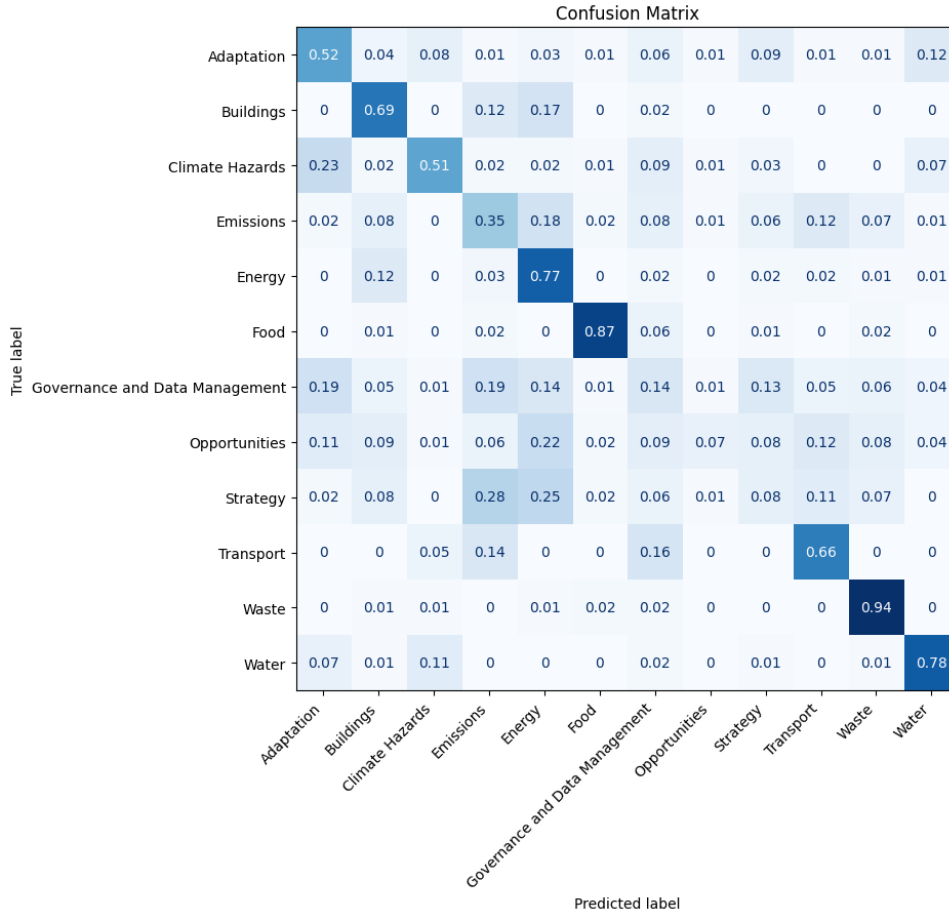


Figure 3: Confusion matrix of GPT-4o-mini predictions on *ClimaTOPIC* dataset.

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
Climate TCFD rec.(Bingler et al., 2023)	16.5(12.9-20.3)	56.5(50.8-61.9)	69.4(64.0-74.4)	68.5(63.0-73.5)	47.5(42.2-53.0)	43.2(37.5-48.0)	50.2(44.4-55.2)	-

Table 10: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers.

category, which accounts for 47.4% of the total errors.

In our sampled errors, we found that 15/30 misclassifications are genuine errors, with 4/30 stemming from the model’s inability to correctly identify strategic actions or cases where the strategic nature of the text is implied but not explicitly stated. The use of a larger model (GPT-4) mitigated this issue to some extent, reducing the rate of strategy-related errors by 75%.

Furthermore, there is notable confusion between the “Strategy” and “Risk” categories. This can be explained by the by the guidelines. The “strategy” label includes descriptions of the climate-related risks, while “risk” includes more specifically risk management, which therefore exclude descriptions

of the risks. Moreover, it seems that the “strategy” label is used as “default” label for vague statement mentioning climate-related topics (2/30). Example C.32 and C.33 are both classified as “strategy” in the gold standard, yet they do not describe climate-related risks and impacts on the company. Overall, we believe that there is a significant overlap between the strategy and risk labels, as describing the risk and describing how the risk is managed are often discussed in the same sentences.

We found 5 examples of texts not focusing on climate-related disclosure but more broadly about sustainability. These examples seem to indicate that this should be considered as related to climate-disclosure. However, even if it mentions sustainability, Example C.31 is mainly about plastic pol-

lution, which we considered as an annotation error.

We also found the issues mentioned on other task: mentioning energy is classified as not climate-related by gpt-4o-mini but systematically climate-related in the dataset (2/30).

Example C.31 (governance). Report 2019 (SR 2019), as requested by the shareholders. The hardcopies will be delivered once they are made available to the Company. * Nevertheless, we hope you would consider the environment before you decide to print the above reports or request for the printed copy of the IR 2019, GFR 2019 and SR 2019. The environmental concerns like global warming, deforestation, climate change and many more affect every human, animal and nation on this planet.

Example C.32 (strategy). The Sustainability Report is based upon the internationally recognized Global Reporting Initiative (GRI) Standards. Our reporting is also guided by the Sustainability Accounting Standards Board (SASB) and the Financial Stability Board’s Task Force on Climate-related Financial Disclosures’ (TCFD) recommendations.

Example C.33 (strategy). The progress toward all Green Company targets is tracked through an environmental data reporting system and is disclosed in detail in our annual Green Company Report, available on our corporate website as of spring 2021. ! > ADIDAS-GROUP.COMSENVIRONMENTAL-APPROACH Own operations: Progress toward 2020 targets

Example C.34 (governance). Report 2019 (SR 2019), as requested by the shareholders. The hardcopies will be delivered once they are made available to the Company. * Nevertheless, we hope you would consider the environment before you decide

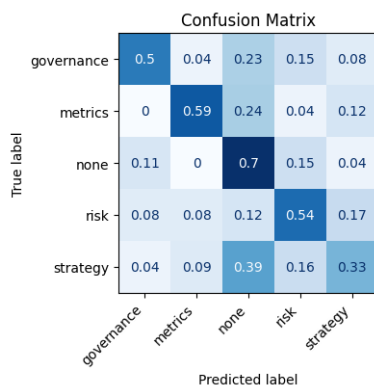


Figure 4: Confusion matrix of GPT-4o-mini predictions on *Climate TCFD recommendations* dataset.

to print the above reports or request for the printed copy of the IR 2019, GFR 2019 and SR 2019. The environmental concerns like global warming, deforestation, climate change and many more affect every human, animal and nation on this planet.

C.4 Detecting Environmental, Social, and Governance Disclosure

Task Description This is a particular case of subtopic classification. The topics are “Environment”, “Social”, and “Governance” categories.

Experiment The only openly available datasets are *ESGBERT E* (Environmental) *ESGBERT S* (Social) and *ESGBERT G* (Governance) datasets (Schimanski et al., 2023b). We evaluated this dataset with the setting described in section B and reported the results in Table 11. Our pre-processing pipeline slightly impacted the dataset when removing duplicates.

Result Analysis In our experiments, we found that for the *ESGBERT E, S and G* datasets, the best performing model is DistilRoBERTa. As we reported the macro average, we also computed the binary F1-score 95%, 88%, 76% (respectively E, S and G), which are close to the performance reported in the original study: 93%, 92%, 79%. This shows that fine-tuned models can identify text relevant to those E, S and G categories. *However, we also see that the TF-IDF baseline performances is over 80% for each dataset, showing that those thematic have largely different vocabulary, making them easily identifiable.* Finally, the zero-shot approaches also performed well on those datasets, with performances above 78% for GPT-4o-mini. Showing that zero-shot models can also identify those categories. Interestingly, the only dataset where the fine-tuned model performed significantly better than GPT-4o-mini is the “Social” dataset.

Moreover, Schimanski et al. (2023b) measured a high inter-annotator agreement (Fleiss’ κ of more than 86%). Overall, this shows that those labels are quite easy to distinguish.

C.4.1 Error analysis

ESGBERT E We are only reviewing errors from *ESGBERT E* as it is the most relevant to our topic. GPT-4o-mini made only 8 errors. 6 FN and 2 FP. 2/6 FN are text mentioning the energy industry (Example C.35) but not directly talking about the environment. 1/6 FN is truncated, 1/6 FN is out-of-context, which makes them difficult to classify.

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
ESGBERT E(Schimanski et al., 2023b)	42.6(35.1-49.3)	88.1(82.5-92.8)	95.8(92.4-98.7)	95.8(92.3-98.7)	95.0(91.3-98.2)	95.1(91.2-98.2)	94.6(90.8-97.6)	93.19 ^{1,e}
ESGBERT G(Schimanski et al., 2023b)	53.5(46.4-60.6)	80.2(73.5-86.1)	78.2(71.7-84.5)	83.7(77.0-89.7)	83.5(77.4-88.8)	75.6(69.0-81.3)	62.6(55.3-69.0)	78.86 ^{1,e}
ESGBERT S(Schimanski et al., 2023b)	54.4(47.0-60.8)	82.6(76.8-88.1)	88.3(83.7-92.6)	89.6(85.2-93.7)	78.1(72.4-84.1)	77.9(71.5-83.4)	73.6(67.3-79.8)	91.90 ^{1,e}

Table 11: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *l*. binary F1-score, *e*. ESGBERT

We also found 1/6 FN mislabeled example talking about the "office environment" (Example C.36). Overall, the errors stem from statement that are systematically classified as related to the environment such as anything related to energy, even if the statement is not explicitly talking about the environment. For the 2 FP, 1/2 is related to the use of "sustainable" in an economic context, and 1/2 describes the impact of floods on mining operations which was classified by GPT-4o-mini as "Environment" but not in the gold standard.

Example C.35 (Environment). e Natural gas liquids for Russia are included in crude oil.

Example C.36 (Environment). The pandemic may also have long-term effects on the nature of the office environment and remote working, which may result in increased costs and present operational and workplace culture challenges that may also adversely affect our business.

C.5 In-depth Disclosure: Climate Risk Classification

Task Description Given an input sentence or a paragraph, output "opportunity" or "risk" label.

Experiment We collected climateBERT’s *climate sentiment* dataset (Bingler et al., 2023). We evaluated this dataset with the setting described in section B and reported the results in Table 12. Our pre-processing did not alter this dataset.

Result Analysis In our experiments, as reported in Table 12, on *climate sentiment*, the best performing solutions are the fine-tuned transformers, with Longformer reaching a macro F1-score of 79.9%. The performance reported in the original study was similar, with a weighted F1-score of 83.8% (Bingler et al., 2023), within the confidence interval of our best performing model. The TF-IDF baseline is performing lower but still good, with a macro F1-score of 69%. This performance shows that the vocabulary for risk and opportunity is quite characteristic and identifiable. The zero-shot approaches’ performances varied from 70.5% for Llama 3.1

70B to 77.3% for GPT-4o-mini approaching the performance of finetuned models.

As shown in the confusion matrix Figure 5, only 4 out of the 68 errors do not include the label neutral (5 out of 68 for distilRoBERTa). This shows that discriminating between polar opposite is easy, but the edge cases can be more challenging. If the majority class is the neutral class, then the performances are not representative of the actual performance of the model. This is further confirmed by the relatively lower inter-annotator agreement compared to previous tasks Friederich et al. (2021) reported a low IAA ($\alpha = 0.20$) and Bingler et al. (2023) reported a moderate IAA (Krippendorff’s $\alpha = 0.61$).

C.5.1 Error analysis

climate sentiment As shown in the confusion matrix displayed in Figure 5, most of the errors are on texts originally labeled as "neutral" (65% of errors). And the model tends to preferably output the "opportunity" label (43% of errors) reducing its precision.

In the sampled errors, we identified several patterns of misclassifications. 2/30 instances involved discussions of corporate ambitions (Example C.37) that were labeled as "opportunity" by GPT-4o-mini, but categorized as "neutral" in the gold standard. These texts are not specifically opportunities, but the mention of "ambitious targets" could be considered "positive" in the guidelines due to: "2) and/or about positive impact of an entity’s activities on the society/environment 3) and/or associates specific positive adjectives to the anticipated, past or present developments and topics covered". 3/30 sampled errors contained discussions of both risks and opportunities (Example C.39), highlighting the lack of clarity in the annotation guidelines for such mixed content. Interestingly, 2/30 examples referenced indirect risks (Examples C.40 and C.41), which were labeled as "neutral" in the gold standard, yet identified as "risk" by the model. Lastly, we found 14/30 instances that were likely mis-

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
climate sentiment(Bingler et al., 2023)	35.9(30.6-41.1)	69.0(63.4-74.4)	79.9(74.8-84.4)	77.7(72.6-82.2)	77.3(72.4-82.2)	71.4(65.8-76.4)	70.5(64.7-75.5)	83.8 ^{3,c}

Table 12: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. 3. weighted average F1-score, c. climateBERT.

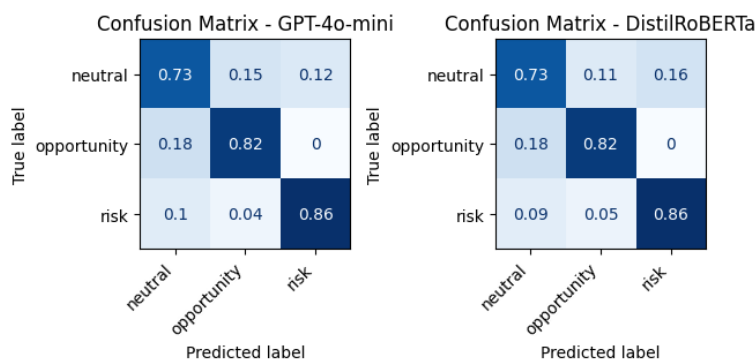


Figure 5: Confusion matrix of GPT-4o-mini predictions on ClimateBERT’s *climate sentiment* dataset.

labeled. We found 5/30 examples of texts that describe a risk-related structure (e.g. Examples C.42), which can reasonably be classified as neutral. While the dataset’s labels are generally well-defined, these observations suggest that further refinement of definitions for certain edge cases could help improve consistency in classification.

Example C.37 (neutral). TD recently launched a bold and ambitious climate action plan to address the challenges of climate change. This includes a target to achieve net-zero greenhouse gas emissions in our operations and financing activities by 2050. We backed this commitment with the creation of a new Sustainable Finance and Corporate Transitions Group to support clients around the world, and an Environmental, Social and Governance (ESG) Centre of Expertise to participate in the global efforts required to deliver on this long-term target.

Example C.38 (neutral). To exclude the least energy efficient hydrocarbons and those that pose the greatest threat to the environment, because these are incompatible with the goal of combating climate change and they represent an economic risk for investors. This means turning down projects and companies that do the majority of their business in: o Oil sands production, o Oil extracted from the Arctic region (off-shore and on-shore production), o Shale gas or oil production involving excessive flaring or venting, o Infrastructure projects mainly intended for schemes covered by the exclusion criteria set out above, Credit Agricole S.A. is

committed to offsetting the Group’s entire direct carbon footprint until 2040 via the

Example C.39 (risk). Sustainability aspects and climate-related risks and opportunities are integrated into the Group-wide risk management process at Daimler. They are understood to be conditions, events, or developments involving environmental, social or governance factors (ESG), the occurrence of which may have an actual or potential impact on the Daimler Group’s profitability, cash flows and financial position, as well as on its reputation. ESG-related risks and opportunities that are very likely to have a serious negative impact on non-financial aspects in accordance with the CSR Directive Implementation Act (CSR-RUG) can be found in the respective categories of the Risk and Opportunity Report according to their cause. Furthermore, Daimler follows the recommendations of the Task Force on Climate-related Financial Disclosures (TCFD) with regard to climate-related risks and opportunities.

Example C.40 (neutral). In responding to Resolution 6.1 we have: z approved and published an Energy Policy consisting of financing of fossil-fuel-related activities, encompassing thermal coal, upstream oil and upstream gas, fossil-fuelled power generation, and renewable and embedded energy solutions.

Example C.41 (neutral). We also look at the way banks facilitate financing by others, for example by arranging the issue of green bonds. For each

of these categories we look at the bank’s current lending, historical trends and lending targets. There are also ‘no go projects’. We won’t invest in any bank which lends to an Adani Carmichael coal mine. Internationally we won’t invest in any bank which lends to a Keystone XL pipeline transporting oil from the tar sands of Canada.

Example C.42 (risk). Our Emerging Risk Committee, constituted by our executive directors, discusses various risks that may affect our business, and includes environmental considerations such as climate change as a standing agenda item. In addition, Newton’s Operating Committee is responsible for the management of how our business is run, and also considers relevant risks. Both committees ultimately report to the Executive Committee, with the outputs from the Emerging Risk Committee being reviewed at the Board Risk Committee.

C.6 Green Claim Detection

Task Description Given an input sentence or a paragraph, output a binary label, “green claim” or “not green claim”.

Experiment We collected the datasets on *Environmental Claims* (Stammach et al., 2023), and on *Green Claims* (Woloszyn et al., 2022) and reproduced the experiments with the setting described in section B and reported the results in Table 13. Our pre-processing pipeline altered the *Green Claims* dataset when removing URLs that are important in tweets.

Results Analysis As shown in Table 13, the best performing models are the fine-tuned transformers reaching F1-scores up to 91.2% on *Environmental Claims* and 97.2% on *Green Claims*. The performances of the zero-shot models are slightly worse by around 5%. Finally, the TF-IDF baseline underperformed the other approaches but still reached macro F1-scores above 80%.

Stammach et al. (2023) reported a binary F1-score of 84.9%. In our experiments, GPT-4o-mini reached a binary F1-score of 81% and DistilRoBERTa reached 87%.

Similarly, Woloszyn et al. (2022) reported a binary F1-score of 92.08%. In our experiments, GPT-4o-mini reached a binary F1-score of 89% and DistilRoBERTa reached 98%. When accounting for the confidence interval (94.84%-100%), our fine-tuned DistilRoBERTa outperformed their best model (fine-tuned RoBERTa). This might be explained by our cleaning process (e.g. removing

URLs) that removed noise from the original dataset, but it could also be due to our split. (Woloszyn et al., 2022) showed that fine-tuned models trained on their dataset are prone to adversarial attacks, which our cleaning process might have exacerbated.

C.6.1 Error analysis

Environmental Claims We found that most of the errors are FP (86%). Most of the errors sampled (9/14) are actual errors by GPT-4o-mini. There are only 4 FN. 3/4 FN are implicit claims. They are descriptions of environment-related actions (Example C.43) but they do not explicitly suggest that the company has a positive environmental impact. We also found 2 examples of implicit claims in the FP (Example C.44). *In the gold standard, implicit claims are labeled sometime positively and sometime negatively. This shows an ambiguity in the annotation guidelines for implicit or indirect claims.*

Most of the sampled FP (7/10) are statements out-of-context that require the annotator to make assumptions. Such statements are well-described in the definition and should be classified as negatives, which makes them actual errors from the model.

Example C.43 (Environmental Claim). We have developed a Climate Change Strategy Roadmap outlining our climate change governance framework.

Example C.44 (Not Environmental Claim). In terms of initiatives regarding energy issues, the AGC Group works to reduce the energy involved in its production activities.

Example C.45 (Not Environmental Claim). This reduces KLP’s climate risk and makes it clear that our work has an effect on the climate.

Example C.46 (Not Environmental Claim). Obviously, it’s without a doubt, the environment continues to improve, and that’s all good.

Green Claims Due to the size of the dataset and the high model performance, we observed only 6 classification errors: 3 FN and 3 FP.

When reviewing the errors, we found that 2 FP (e.g. Example C.47 with “Sea buckthorn extracts”) and 1 FN (Example C.49 with “Coconut Oil”) fall into the “natural claim” category - suggesting that a product is better, healthier, or has properties thanks to a natural ingredient. It is not clear how these claims should be annotated. In the guidelines, they

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
Environmental Claims(Stammbach et al., 2023)	42.8(37.4-48.4)	80.5(74.8-85.5)	90.5(86.5-94.0)	91.2(86.9-94.6)	86.8(82.0-90.9)	81.4(76.0-86.1)	76.6(71.1-81.9)	84.9 ^{1,r}
Green Claims(Woloszyn et al., 2022)	46.1(34.9-56.7)	86.1(77.2-93.3)	94.5(88.7-98.7)	97.2(93.0-100.0)	91.5(83.9-97.4)	90.0(82.3-96.9)	86.8(78.9-93.5)	92.08 ^{1,r}

Table 13: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *l*. binary F1-score, *r*. RoBERTa

could be considered “nature-friendly messages that target the needs and desires of environmentally concerned stakeholders” but they could also be considered as “Claims that address issues not related to the environment, such as health or equality, [which] are not considered green claims”. (Woloszyn et al., 2022) reported that: “natural ingredients” and “naturally derived” were the most frequent bigrams in “Explicit Green Claims”. We found that the tweets containing these bigrams are “natural claim”, therefore we considered that “natural claims” are part of “green claims”. This implies that the 2 FP are mislabeled examples, and the FN and error by GPT-4o-mini.

The 2 other FN, are out-of-context, generic statements that do not refer to any specific product (Example C.48). Whether such statements should be classified as “green claims” is debatable. If we consider that any communication on environmental topics is an attempt to portray the company in a more environmentally friendly light, then these statements could be interpreted as green claims. However, this approach is overly simplistic and risks categorizing any environmental reference as a green claim, which would be an over-extension of the definition.

The remaining FP represents an actual misclassification by the model.

Example C.47 (Not Green). He needs protection from harsh weather too. Sea buckthorn extracts fortify + invigorate for smooth, energized skin: <URL> <URL>

Example C.48 (Green Claim). Environmental solutions come in all shapes and sizes.

Example C.49 (Green Claim). TELL US: what do you love the most about #WholeBlends Smoothing with #Coconut Oil and #Cocoa Butter Extracts? #haircare #coconutoil <URL>

C.7 Green Claim Characteristics

Task Description Given an input sentence or a paragraph labeled as green claim, output a more fine-grained characterization of the claim. This is

a multi-label classification task; the labels can be about the form (e.g. specificity) or the substance (e.g. action, targets, facts).

Experiment All the previously mentioned datasets are available. We collected them and reproduced the experiments with the setting describe in section B and reported the results in Table 14. *Implicit/Explicit Green Claims* (Woloszyn et al., 2022) was impacted by our pre-processing pipeline when we removed URLs.

Results Analysis As shown in Table 14, we observed the same patterns for each dataset. The best-performing models are fine-tuned transformers reaching performances above 77%. The zero-shot approaches underperformed the fine-tuned models by 1 to 5%. The TF-IDF baseline performed significantly better than random, underperforming the best models by between 2 to 10% only. It is important to note that the difference in performance with the TF-IDF baseline is significant only for DistilRoBERTa on *Commitments And Actions* and for Longformer on *Net-zero/Reduction*. This suggests that the characteristics mentioned, both in terms of form and content, are expressed with a specific vocabulary that can be used as a decent predictor. On *Implicit/Explicit Green Claims* and *Net-zero/Reduction*, GPT-4o-mini reached similar performances as the fine-tuned models. On *Climate Specificity* and *ESGBERT action500* GPT-4o-mini underperformed both the TF-IDF baseline and the fine-tuned models. On *Commitments and Actions*, the fine-tuned models significantly outperform the zero-shot approaches.

The performance reached by fine-tuned models in our experiments is similar to those reached in the original studies.

C.7.1 Error analysis

Commitments And Actions The majority of all errors are FP (77%). Another ambiguous aspect is the description of company philosophy/values/mindset and the description of governance structures. We found examples of philoso-

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
Commitments&Actions(Bingler et al., 2023)	48.5(42.9-53.9)	72.7(67.4-78.2)	76.7(71.2-81.5)	81.9(77.1-86.5)	67.2(61.6-72.4)	64.6(59.4-69.7)	50.7(45.1-56.2)	81 ^{3,c}
Climate Specificity(Bingler et al., 2023)	48.9(43.6-54.3)	72.4(67.6-77.5)	77.3(72.4-81.9)	77.5(72.1-82.2)	71.6(66.6-76.3)	76.5(71.5-80.8)	68.0(62.8-73.2)	77 ^{3,c}
ESGBERT action500(Schimanski et al., 2023b)	44.4(29.9-58.5)	82.5(71.1-93.1)	85.9(75.4-95.8)	89.1(80.2-97.8)	76.0(62.2-87.4)	67.0(53.3-79.7)	63.0(47.7-76.1)	-
Implicit/Explicit Green Claims(Woloszyn et al., 2022)	37.2(26.3-46.1)	70.3(56.5-82.0)	63.0(48.8-74.9)	81.2(68.1-91.8)	81.6(71.2-90.4)	68.5(54.7-81.0)	75.7(63.4-86.0)	81.45 ^{2,r}
Net-Zero/Reduction(Schimanski et al., 2023a)	29.8(24.9-34.8)	94.8(92.2-97.1)	97.8(96.0-99.1)	97.7(95.8-99.3)	97.3(95.5-99.0)	93.7(90.7-96.2)	95.6(93.3-97.8)	98.7 ^{1,c}

Table 14: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. 1. binary F1-score, 2. average F1-score (macro/micro is not specified), 3. weighted average F1-score, c. climateBERT, r. RoBERTa

phy with the positive and negative labels (Examples C.51, C.52) and examples of governance in both (Examples C.53, C.52). Those kinds of errors represent 3/10 sampled FN and 7/10 of FP. Indeed expressing company values could be considered as commitments toward these values or a simple description of them; similarly, describing a governance structure could be considered as an action (setting up the governance structure) or a simple description. This ambiguity on what constitutes a commitment and an action is also reflected in the moderate inter-annotator agreement among human annotators (Krippendorff’s α of 0.5354).

Additionally, this task is not climate-specific, and therefore the guidelines include both “business or climate actions”. GPT-4o-mini understood the prompt as only “climate actions”. 2 FN are due to GPT-4o-mini being anchored to a word in the prompt. Using a larger model or updating the prompt might resolve this issue. There is also 1 FN that was mislabeled (Example C.54), as the text describes a process and does not contain an action or a commitment. The remaining FN (4/10) are actual errors.

Of the 3 remaining FP sampled, 2 are mislabeled, such as Example C.50, which describes an action about recycling but is labeled as “No” in the gold standard - and as an “action” by GPT-4o-mini.

Example C.50 (No). 1 Recycling: To meet the environmental policies of major countries, we operate a recycling process utilizing non-reusable dead batteries and scraps generated during battery production. We have a strategic cooperation relationship with major partners for realizing a closed-loop. Through this system, we extract raw materials of batteries such as nickel, cobalt, and lithium by crushing and dissolving battery waste or scraps.

Example C.51 (Yes). In our view, true sustainable investing cannot be achieved by simply voting a proxy, adding a director of sustainability or even divesting from an asset class. Because traditional

models of finance and investing often fail to appropriately integrate sustainability issues, we’ve had to build it into our thinking from the ground up. *It requires integration across our products, across our product teams and across our entire organization.*

Example C.52 (No). Climate risk has emerged as one of the top environmental risks for the Bank. This includes physical risks related to the chronic and acute physical impacts of climate change (e.g., shifts in climate norms, and extreme weather events such as hurricanes, wildfires and floods), and transition risks associated with the global transition to a low-carbon economy (e.g., climate-related policy actions and litigation claims, technological innovations, and shifts in supply and demand for certain commodities, products and services). Both physical and transition risks could result in strategic, credit, operational, legal, and reputational risks for the Bank and its clients in climate sensitive sectors. *TD supports Canada’s objectives to meet the goals of the Paris Agreement and recognizes the Bank’s responsibility to contribute by integrating climate considerations across its business.* The Bank continues to monitor industry and regulatory developments and assess the potential impacts of climate change and related risks on its operations, lending portfolios, investments, and businesses.

Example C.53 (Yes). Our sustainability efforts are underpinned by strong corporate governance, which we continue to reinforce in line with recommendations of the Malaysian Code on Corporate Governance 2017 (MCCG 2017). This saw us set up a new Board Risk Committee to further enhance our risk oversight, adding to the existing Board Audit Committee and Nomination and Remuneration Committee. The Board as a whole comprises four Independent Non-Executive Directors (INEDs), who make up 50% of its composition, two of whom are foreign Directors. On a related note, I represented the Group in signing an Integrity Pact with the Malaysian Anti-Corruption Commission

(MACC), underlining our commitment to observing integrity in all our dealings with stakeholders.

Example C.54 (Yes). D+ This client does not (yet) meet Rabobank’s sustainability policy on one or more points or has not responded adequately to key questions. Specific agreements are made about a possible solution and timelines are established. Once the customer meets the sustainability policy of Rabobank, it is classified in category A, B or C.

Example C.55 (No). Governance structure A robust governance structure ensures timely and direct execution of programs that drive the achievement of our set goals for 2020 as well as of our new set of targets for 2025. The head of Sustainability is responsible for the development, coordination and execution of our sustainability strategy and reports to the member of the Executive Board responsible for Global Operations. He or she also leads the sustainability Sponsor Board, which is composed of senior representatives from Global Brands, Global Operations, Digital, Sales, Finance, Corporate Communication, and other relevant functions across the company. The Sponsor Board ensures cross-functional alignment, transparent end-to-end management and execution of agreed-upon sustainability goals within their functions. This includes reviewing and signing-off on policies as required. We also maintain a separate compliance function which is operated as the Social & Environmental Affairs Team (SEA) to evaluate supplier-facing social and environmental compliance performance and human rights impacts, reporting, through the General Counsel, to the CEO.

This dataset would require guidelines that are more detailed and specific for multiple ambiguities (that are not clear neither in the guidelines nor in throughout the annotations) :

- Setting a governance structure or process an action ?
- Is describing a existing process (industrial, governance, ...) an action ?
- Is saying that the company want to implement sustainability across all their products an action/commitment ?

ESGBERT’s Actions500 GPT-4o-mini produced only 10 errors, 9 of which were FP. Since this dataset is not part of a publication, we lack access to detailed annotation guidelines, leaving certain

edge cases, such as statements about commitments, company philosophy, or exploration of potential solutions, open to interpretation. Notably, 13/19 of these errors involve cases that are debatable, suggesting that clearer guidelines could improve consistency.

Climate Specificity In GPT-4o-mini predictions for *Climate Specificity*, most of the errors are FP (93%). There are actually only 6 FN.

When investigating the FN, we found 2 examples of part of the legend of a figure such as Example C.56. In the guidelines, the legend “provides firm-specific detailed explanations to enable readers to better understand the overall information reported” and is therefore considered “specific”. We also found 2 examples which were mislabeled as the statement applies to the whole industry, which is considered “non-specific” in the guidelines. We classified Examples C.57 as out-of-context, as they mention "A similar approach could be used" which we argue is not specific. But more context could make it specific.

Out of the sampled FP, we found 4/10 debatable examples. *This is due to the granularity of the specificity.* A paragraph might contain a list of specific “Key Priorities”. However, as the priorities are strategic axes, they are not specific about actions or targets. For instance, Example C.58 is specific about using "red lines" but does not describe specifically what the "red lines" are. Similarly to *Commitments and Actions*, the statements about governance can be precise governance points - such as the role of the CEO in overseeing the climate-related issue - but not specific about actions or targets. This is the case for 2 of the sampled errors.

The dataset contains annotation for identification of “specific actions/targets specific to the company”. This task is particularly difficult, as it needs to be precisely defined and even with a precise definition, we believe that some statements remain ambiguous due to the granularity of the specificity. This is further confirmed by the low IAA (Krippendorff’s α of 0.1703)

Example C.56 (specific). [1] Data calculated on the same reporting perimeter as 2018, excluding Abertis Group. [2] Data on the diesel consumed by the gensets in Chile in 2018 are not available. [3] Figure updated following a consolidation subsequent to the close of the Integrated Report 2018 on ETC data. [4] Data updated following a restatement of the income statement 2018. [5] Location-based

emissions.

Example C.57 (specific). A similar approach could be used for allocating emissions in the fossil fuel electricity supply chain between coal miners, transporters and generators. We don't invest in fossil fuel companies, but those investors who do should account properly for their role in the production of dangerous emissions from burning fossil fuels.

Example C.58 (not-specific). Sustainable strategy 'red lines' For our sustainable strategy range, we incorporate a series of proprietary 'red lines' in order to ensure the poorest- performing companies from an ESG perspective are not eligible for investment.

Implicit/Explicit Green Claims As the dataset is small, there are only 12 errors. As show in figure 6, the most frequent error, is a statement without claims classified as an "implicit claim".

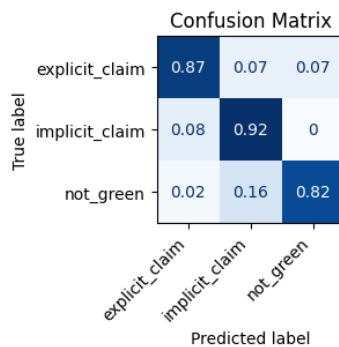


Figure 6: Confusion matrix of GPT-4o-mini predictions on *Explicit/Implicit Green Claims* dataset.

As discussed in Section C.6, it is not clear if green claims include "natural claim" - suggesting that a product is better, healthier, or has properties thanks to a natural ingredient. Among the 12 errors, 5 contain "natural claims" (Example C.60). We also found 2 errors that are debatable, and 2 annotation errors. For instance, the annotator classified Example C.59 as "not green". They most likely understood "Irish Spring" as the season instead of the soap brand. On the contrary, GPT-4o-mini correctly classified it as "implicit green claim". This shows that LLMs can have broader general knowledge than the original annotators. *The distinction between explicit and implicit is highly relevant, as it define straightforward cases from ambiguous cases which are harder to classify.* Despite those errors, overall both fine-tuned models and the zero-shot approaches reached good performances. This is

aligned with the good IAA reported in the original study (Woloszyn et al., 2022) (Krippendorff $\alpha = 0.8223$).

Example C.59 (not green). Irish Spring is green in more ways than one... <URL>

Example C.60 (not green). Give your skin a mid-day refresh with the Yes To Grapefruit Unicorn Brightening Mist! It's naturally packed with vitamin C to help give your skin a boost of GLOW! Thanks, @POPSUGAR <URL>

Example C.61 (green). A healthy lifestyle becomes ever more important for many of us - also within our beauty routine. Stay tuned for our new permanent hair color featuring natural ingredients such as soy protein, oat milk and argan oil! Available from February. #headsap #WhatsComingNext #ProductNews <URL>

Net-zero/Reduction On this dataset, the model failed only on 8 instances as shown in Figure 7. As the performances of models and TF-IDF baseline are really high, we also experimented with using only a few keywords. Using ["net", "zero", "neutrality"] for the "net-zero" class and ["reduction", "reduce"] for the "reduction" class, we reached an F1-score of 78.03%.

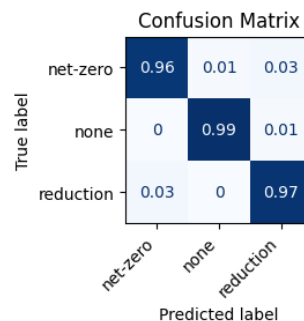


Figure 7: Confusion matrix of GPT-4o-mini predictions on *Net-zero/Reduction* dataset.

Among the errors, 1 is a simple statement classified as targets in the gold standard (Example C.62). Another example is labeled "net-zero" while dealing with a reduction of 95% (Example C.64), raising the question: is 95% reduction the same as net-zero? Finally, half of the errors are statements with both "reduction" and "net-zero" targets (Example C.63). Overall, we see that both fine-tuned models and the zero-shot approaches are proficient at identifying targets. However, we point out that without addressing the labeling error mentioned, we can't conclude on the performance differences. While

the task appears straightforward for models, annotators do not achieve perfect inter-annotator agreement, despite a high score (Cohen’s $\kappa = 0.931$). This indicates that even seemingly simple tasks can still be subject to occasional disagreement among annotators.

Example C.62 (net-zero). Based on our track record and plans up to 2030 we are confident that we are well-placed to make significant progress beyond 2030 and up to 2050. (Pg. 27)

Example C.63 (reduction). Reduce GHG emissions (Scopes 1+2) (on a market basis*1) by 55% by FY2030 and to zero by FY2050 (compared to FY2017)

Example C.64 (net-zero). Amsterdam aims for a 95% emissions reduction by 2050 (reference year 1990). Unclear how this becomes climate neutrality though.

C.8 Green Stance Detection

Task Description Given two input sentences or paragraphs, one labeled as the claim and one as the evidence, predict the stance between the two: supports, refutes or neutral. Some studies fix the claim and only vary the evidence (e.g. the claim is always *Climate change poses a severe threat*), training a model to predict the stance of the evidence in respect to the fixed claim. Other studies train a model to predict the relation between varying claims/evidences.

Experiment We collected *ClimateFEVER* (Diggelmann et al., 2020), *climateStance* (Vaid et al., 2022), *GWSD* (Luo et al., 2020) and *LobbyMap* (Morio and Manning, 2023) and reproduced the experiments with the setting described in section B and reported the results in Table 15. For *ClimateFEVER* the dataset is designed to predict the label of claim-evidence pair relation and then aggregate the predictions to infer if the claim is supported or not. Therefore, we split the tasks into 3. The first is to predict the claim label only, without the evidence (*ClimateFEVER claim (our split)*). The second task is to predict the claim-evidence label (*ClimateFEVER evidence (our split)*). Finally, the third task is given the prediction of the claim-evidence pair, predict the claim label (*ClimateFEVER claim (our split, aggregated)*). We also reported the performance for a train-test split similar to climabench (Spokoyny et al., 2023). For *Lobbymap*, we evaluated each

intermediary task (Page, Query, Stance). Our pre-processing impacted *ClimateFEVER* and *Lobbymap* as we resampled the datasets and removed longer texts. It also altered *climateStance*, as we removed URLs which are important in tweets.

Results Analysis When focusing on the results for *ClimateFEVER claim (our split, aggregated)*, *ClimateFEVER evidence (our split)*, *ClimateStance*, *GWSD* and *LobbyMap* datasets; our experiments results show performances below 75% for the best-performing models. This shows that those tasks are challenging for models. For most datasets, the best-performing approaches are the fine-tuned models. However, for *ClimateFEVER*, it is the zero-shot approach.

When using a split similar to Spokoyny et al. (2023) on *climateFEVER evidence (climabench split)*, distilRoBERTa reached an F1-score of 59.9% (comparable to theirs, with 62.7% for (Spokoyny et al., 2023)). However, as mentioned in Appendix A, the split provided in their benchmark contains partial contamination (the same claim is both in the train and test splits, but with different evidences). When using our split, the performance of the best-performing fine-tuned transformer drops to 52.7%. On *ClimateFEVER claim* - classify if the claim is supported by evidences - we first evaluated if the claim contains a predictive power in itself. Both fine-tuned models and zero-shot approaches struggle on this task. This is likely due to the way the labels are constructed: the claim label is aggregated on the claim-evidence labels. Therefore, the claim’s label is dependent on the evidence collected previously and not on its inherent factuality. When using aggregated claim-evidence prediction to infer the claim’s label, the performance increases to 41.3%. Still, this is far from the performances reported in other papers (Wang et al., 2021; Vaghefi et al., 2022; Xiang and Fujii, 2023; Webersinke et al., 2022; Vaid et al., 2022) reaching performances above 80%. However, they are not comparable as they removed claims with the "DISPUTED" label.

We achieved performance levels comparable to those reported in the original studies across all other datasets except for *LobbyMap*. This discrepancy can be attributed to the reduced size of our training dataset, which was limited to 10,000 samples. Given the complexity of the task, a larger training set or extended training duration could po-

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
climateF: claim (our split)(Diggelmann et al., 2020)	15.4(10.4-20.7)	35.0(26.4-43.0)	31.9(24.7-38.9)	32.5(24.9-39.6)	19.3(13.2-25.0)	30.0(21.9-37.2)	26.3(19.8-32.8)	80.7 ^t
climateF: claim (agg.) (Diggelmann et al., 2020)	-	-	42.5(34.2-49.9)	41.3(32.8-49.9)	50.4(42.0-58.2)	-	-	60.10 ^{z,s,h}
climateF: claim (climabench split, agg.)(Diggelmann et al., 2020)	-	-	11.9(10.7-12.8)	48.8(45.3-52.4)	-	-	-	-
climateFEVER evidence(Diggelmann et al., 2020)	28.4(25.2-31.4)	46.2(42.2-50.2)	52.7(48.5-56.7)	51.3(47.3-55.1)	60.0(56.2-63.6)	52.5(48.5-55.9)	63.7(59.7-67.4)	68.03 ^{z,t,h}
climateFEVER evidence (climabench split)(Diggelmann et al., 2020)	30.4(27.4-33.5)	57.0(52.6-60.7)	26.3(25.5-27.1)	59.9(55.5-64.0)	-	-	-	62.68 ^t
climateStance(Vaid et al., 2022)	21.3(17.7-25.1)	49.6(42.9-55.9)	56.6(49.4-63.0)	56.1(47.9-63.2)	56.4(50.4-62.2)	48.5(42.4-54.2)	58.0(51.2-63.8)	59.69 ^r
Global-Warming Stance (GWSD)(Luo et al., 2020)	29.5(23.0-36.0)	59.2(51.5-66.2)	69.3(62.2-75.8)	75.3(68.5-81.6)	69.8(63.1-76.3)	64.4(57.5-71.1)	72.5(65.8-78.8)	73 ^t
LobbyMap (Pages)(Morio and Manning, 2023)	46.3(45.5-47.1)	73.1(72.3-73.9)	71.3(70.4-72.1)	73.6(72.8-74.4)	63.2(62.4-64.0)	62.5(61.7-63.3)	60.5(59.6-61.3)	-
LobbyMap (Query)(Morio and Manning, 2023)	12.4(12.0-12.9)	49.3(47.3-50.9)	57.3(55.0-59.2)	52.1(50.0-54.2)	36.4(34.5-38.0)	27.6(26.0-29.2)	32.7(31.1-34.2)	-
LobbyMap (Stance)(Morio and Manning, 2023)	19.2(17.8-20.5)	43.6(41.7-45.2)	46.7(45.0-48.3)	44.7(42.8-46.5)	30.0(28.5-31.7)	26.9(25.4-28.4)	24.1(22.6-25.6)	-

Table 15: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *d*. Contaminated split, SciBERT, *h*. Human Baseline using annotations, *i*. Filtered Split (by removing disputed claims), *r*. RoBERTa, *t*. BERT

tentially enhance model performance by providing more comprehensive coverage of the data distribution and enabling the model to learn more nuanced patterns.

C.8.1 Error analysis

ClimateFEVER We will not investigate errors on the claim classification task as the performance is really poor and the task is not designed for this.

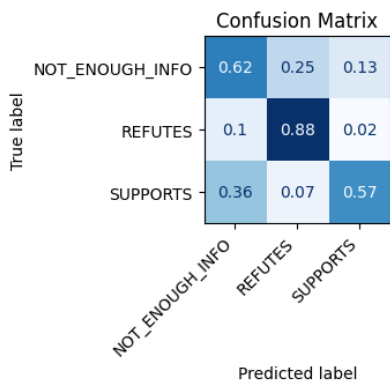


Figure 8: Confusion matrix of GPT-4o-mini predictions on *ClimateFEVER* (evidence) dataset.

On the claim-evidence pair relation classification task, we firstly see that the model rarely mixes ‘SUPPORT’ and ‘REFUTES’ (less than 5% of errors) as reported in Figure 8. Then we see that GPT-4o-mini tend to predict “REFUTES” instead of “NOT ENOUGH INFO”; and “NOT ENOUGH INFO” instead of “SUPPORTS”.

Upon examining the errors, we found that 12/30 are mislabeled in the gold standard. Most of these annotation mistakes (9/12) are evidence that can be classified as SUPPORT/REFUTES but were classified as NOT ENOUGH INFO. This shows that many examples are ambiguous as they are at the frontier of NOT ENOUGH INFO’s label, and this is also aligned with the LLM’s error pattern. we

found that 15/30 errors were genuine misclassifications by GPT-4o-mini, and three instances being potentially debatable. These errors are likely to be mitigated through prompt engineering, for instance, by emphasizing that the evidence was retrieved by a model with potential limitations. Consequently, although the evidence may appear related to the claim, additional verification is required to ensure its relevance. For example, in “He concluded”, “He” is not necessarily the person mentioned in the claim (Example C.65).

As mentioned, we found some errors that are debatable, and Diggelmann et al. (2020) reported a low Cohen’s $\kappa = 0.334$. This shows that the task is difficult even for humans, and there are numerous disagreements. As the individual annotations were provided in the dataset, we also computed the performances of each annotator. We found that the average macro F1-score on the claim-evidence task is 60%, ranging from 54% to 69%. The model is well within the range of the human annotator performances.

Example C.65 (NOT ENOUGH INFO).

Claim: "When you read Phil Jones’ actual words, you see he’s saying there is a warming trend

Evidence: "From this, he concluded that The post-1980 global warming trend from surface thermometers is not credible."

Additionally, we sampled 10 examples labeled “SUPPORTS” predicted as “NOT ENOUGH INFO”. All 10 are claims supporting climate change, and evidence indeed supports the claim. However, the evidence is not sufficient on its own (supporting only a part of the claim or indirectly supporting the claim). We also sampled 10 examples labeled “NOT ENOUGH INFO” predicted as “REFUTES”. 8 claims deny climate change, and 2 are neutral. Except for 2, which are out of context, they all either refute or partly refute (refuting

part of the claim or indirectly refuting the claim). The stance can be interpreted as: “the evidence is sufficient to support/refute the claim” or “the evidence is aligned with the claim and can be part of an argument that supports/refutes the claim”. The annotators seem to struggle with this distinction, while GPT-4o-mini tends to predict “REFUTES” for arguments refuting climate change denial arguments, and “NOT ENOUGH INFO” for arguments supporting climate change arguments.

Climate Stance This dataset is heavily imbalanced toward the “Favor” label. Therefore, most of the errors are on this label. However, as shown in Figure 9 the rate of error is not significantly higher.

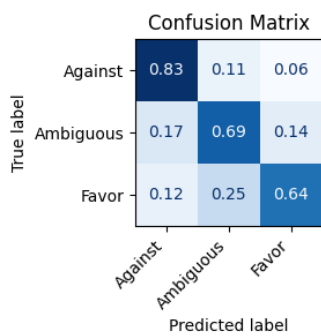


Figure 9: Confusion matrix of GPT-4o-mini predictions on *Climate Stance* dataset.

The definition of the stance provided by Vaid et al. (2022) contains some gray area. For example, the “Against” label includes: “opposition to climate change policies”. Therefore, a statement that includes acknowledging climate change, but denying the actions that are required to limit it, is difficult to classify (Example C.66).

In the sampled errors, 7/30 are statements implicitly expressing concerns about climate change, such as promoting an IPCC conference or making a joke (Example C.67) which are misclassified by GPT-4o-mini. The other sampled errors (23/30) are debatable or mislabeled. They are 11 statements out-of-context and could be interpreted (e.g. Example C.68), or 4 statements clearly mislabeled (e.g. Example C.66 labeled “Favor” in the dataset but correctly classified as “Against” by GPT-4o-mini).

Contrary to other datasets, there are many errors where the model predicts “Against” instead of “Favor”. Therefore, we investigated a sample of 10 errors of this type. We found 6 mislabeled examples (e.g. Example C.69), 1 tweet with an ambiguous stance, and 3 that are actual errors by GPT-4o-mini. This suggests that a lot of those errors are

actually annotation mistakes, most likely caused by the reference to personalities (e.g. “Christopher Monckton”, “Corbyn”) and implicit stance (e.g. “carbon border tax” will damage global efforts to tackle #ClimateChange”) which are not trivial to understand.

Example C.66 (Favor). My point is China uses it devolving status to get an easy ride - while Australia who is making reductions - continues down a silly track of higher energy prices for no real change in outcomes for this so called climate emergency

Example C.67 (Favor). living through climate change, I know what it’s like to be a towns person in an RPG who goes, "there sure are a lot more wandering monsters than there used to be!"

Example C.68 (Favor). But are solutions towards zero emissions as promising as they seem? This well-documented story by @johncarlosoaez contains some sobering numbers: <URL> via @NautilusMag #ClimateChange #ClimateCrisis 2/2

Example C.69 (Favor). Power curve doesn’t exist just like global warming

GWSD In this dataset the “disagrees” label is under-represented. As shown in the confusion matrix (Figure 10), most of the errors involve the label “neutral”. Confusions between “agrees” and “disagrees” represent less than 15% of the errors.

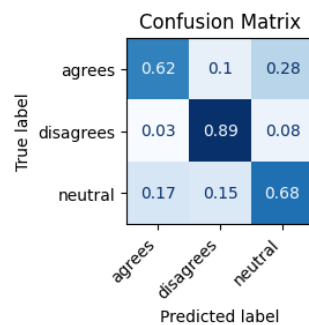


Figure 10: Confusion matrix of GPT-4o-mini predictions on *GWSD* dataset.

With our prompt, the models often identified factual statements as non-opinionated, leading them to incorrectly assign the “neutral” label to statements that clearly conveyed agreement with the target opinion. In multiple examples (10), such as Example C.70, the model correctly recognized that the fact implied the “seriousness” of the situation. However, because it did not contain an explicit opinion, the model classified it as “neutral”. This

pattern accounts for 14 of the 30 sampled errors (with 10 correct reasonings). Interestingly, this type of answer (correct reasoning but wrong label) is inconsistent, in similar situation the model does answer with agrees/disagrees labels for implicit statements.

Additionally, we identified 7 instances of incorrect labeling in the original dataset. One such example, Example C.71), discusses the attention given to global warming, suggesting that it does not match the severity of the issue. While using a prompt that emphasizes implicit opinions did address some of these errors, it did not lead to overall performance improvements. Although it corrected the previously identified mistakes, it introduced new errors in the opposite direction. These newly introduced errors include misclassified examples, such as Example C.72, as well as statements that do not clearly convey the authors’ opinions. *This demonstrates the challenge of dealing with implicit statements, where the inherent ambiguity makes it difficult for the model to assign the correct label.*

Example C.70 (agrees). Global temperatures in 2014 shattered earlier records, making 2014 the hottest year since record-keeping began in 1880.

Example C.71 (neutral). The stark truth is that severe weather events alone will not cause global warming to pop to the top of the national agenda.

Example C.72 (neutral). This summer is seeing record lows for Arctic ice.

LobbyMap Firstly, it is important to note that this dataset is constructed based on LobbyMap.org. While LobbyMap.org is curated by experts, the site does not claim to be exhaustive. As a result, it is possible that the collected stances are not fully comprehensive either. For this dataset, the most important metric, therefore, is the recall.

<i>binary</i>	Precision	Recall	F1-score
Page	50%	94%	65%
<i>macro</i>	Precision	Recall	F1-score
Query	34%	54%	40%
Stance	30%	29%	24%
Stance (relieved)	47%	46%	45%

Table 16: Detailed performances of GPT-4o-mini on the LobbyMap dataset different tasks.

As shown in Table 16, the recall is significantly higher for both the Query and Page tasks. This

is expected, as the dataset is non-exhaustive, leading to lower precision. However, this limitation does not impact the stance task. Since the page and policy query are predefined, the annotators from LobbyMap.org might have overlooked a page containing a stance on a policy. However, given a specific page and policy, there is only one correct stance to identify. In fact, on the Query task, the GPT-4o-mini outputs for each page on average 2.7 queries, while the original labels contain on average 1.6.

As this dataset contains 3 tasks, and that the texts are relatively long, we analyzed the Page and Query task simultaneously. We sampled 10 instances. For the Page task: 3 were TN (true negatives), 1 was a TP (true positive), 1 was a FP. The FP is due to lack of context (Example C.74). The 5 others are texts that actually contain one or more pieces of evidence about the stance of the company. For instance, Example C.73 was not classified as containing a stance in the original dataset; however, the statement contains evidence about the stance of the company on "Energy transition & zero carbon technologies" through the mention of electric vehicles.

For Stance task, it appears that GPT-4o-mini almost never output the “supporting” label. Using the relaxed labels from the original paper, we reach an F1-score of 41%, which is still very low. We also sampled 10 errors. Half of them are labels that are close (“Supporting” and “Strongly Supporting”). We also found 2 instances where the annotators assigned the “not supporting” instead of the “no position or mixed position” label to statements with no position about a policy. The stance is actually expressed in numerical value (between -2 and +2) on LobbyMap.org, but was translated in labels (e.g “supporting (+1)”) in *LobbyMap (Stance)*. The rest of the errors are genuine mistakes.

Example C.73. We’ll show how we’ve revolutionized and electrified some of the most popular, iconic vehicles, helping to shape the future of zero-emissions transportation [No position]

Example C.74. David Maxwell, CEO of east coast gas producer Cooper Energy, described the policy as "draconian" [No position]

In our benchmark, we measure F1-scores for each task individually; however, (Morio and Manning, 2023) proposed custom metrics that we reported in Table 17. We observe that the performances of the Most-Frequent and the Linear Model

	Model	Document			Page Overlap			Strict		
		P	Q	S	P	Q	S	P	Q	S
Morio and Manning (2023)	Most-frequent	46.7	52.6	36.8	51.8	25.6	19.8	41.2	19.6	17.5
	Linear	66.0	61.9	50.3	71.4	44.5	36.1	52.0	31.2	27.0
	BERT-base	71.0	63.5	51.6	73.6	48.1	37.2	50.2	31.9	25.8
	ClimateBERT	71.8	64.0	52.8	74.4	48.9	39.0	50.2	32.2	26.8
	RoBERTa-base	71.6	64.5	53.1	73.8	49.6	38.3	50.4	33.4	26.6
	Longformer-base	73.7	66.9	54.6	75.9	53.0	40.8	52.5	36.1	28.6
	Longformer-large	73.9	68.8	57.3	76.5	55.0	44.1	53.6	38.7	31.5
Ours	GPT-4o-mini	58.6	33.5	39.7	68.6	22.6	23.9	41.0	11.6	14.2
	distilRoBERTa	61.2			-			39.23		
	TF-IDF	63.5	57.4	50.2	65.9	42.6	34.7	39.3	25.5	20.9
	Most Frequent	46.7	52.6	36.8	52.0	25.7	19.8	41.2	19.6	17.5

Table 17: Performance comparison of various models across different evaluation metrics.

were easily reproducible. However, our finetuned distilRoBERTa could not reach the same performances as Morio and Manning (2023)’s finetuned models. Moreover, GPT-4o-mini could not output better results than the TF-IDF/linear baseline.

C.9 Question Answering

Task Description Given an input question and a set of resources (paragraphs or documents), output an answer to the question.

Experiments We collected *ClimaQA* and *ClimaINS* and reproduced the experiments with the setting described in Section B and reported the results in Table 18. Our pre-processing step heavily altered the datasets as they contained many problematic duplicates. We found that 10% *ClimaQA*’s dataset are duplicates with mismatching labels (a given question and its answer would appear with both positive and negative labels). We had to rebuild the dataset from the original CDP questionnaires answered by the companies. This dataset is called *ClimaQA (our split)* in the evaluation. *ClimaINS* also contained many duplicates due to the companies answering similarly each year, and companies answering similarly to the parent-company (reducing the dataset size by 25% and avoiding contamination). Around 2.4% of the remaining answers are duplicates with mismatching labels. We clean the dataset by removing all problematic examples, we name this dataset *ClimaINS (our split)*. We framed *ClimaQA* as a classification task, to fit the settings of our experiments.

Results Analysis Firstly, for *ClimaINS*, in the original paper the authors reported a performance of 84.4%, which we could reproduce with the same settings, even reaching 89.3% with DistilRoBERTa.

However, as described previously, we significantly altered the dataset, removing duplicates and train-test contamination. This resulted in lower performances from fine-tuned models, both Longformer and DistilRoBERTa, as shown in Table 18. With this split, all models are under-performing the TF-IDF baseline. This suggests that while each question induces a different vocabulary, understanding the responses actually makes it harder to identify the source question.

For *ClimaQA*, with the original split the models could not reach more than an F1-score of 50%. This was due to the duplicates with mismatching labels introducing noise. However, when re-constructing the dataset, the fine-tuned models can reach excellent performances, around 89% as shown in Table 18. Contrary to *ClimaINS*, the zero-shot approaches do not under-perform the TF-IDF baseline. Moreover, the gap with the fine-tuned models is smaller (around 9%).

In the original paper, the authors framed *ClimaQA* as a retrieval task, using models to rank the answers. The best model is MiniLM, reaching a MRR (Mean Reciprocal Ranking) of 0.755. We also computed the MRR of the distilRoBERTa finetuned during our classification experiment, as well as a *random* baseline and a *distribution* baseline in Table 19. The *distribution* baseline always ranks the question based on the frequency of the question in the training set. As reported in the table, the finetuned model reached a MRR of only 0.636. We believe that the performance might be improved by using a larger training dataset than ours (limited to 10,000 samples).

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
ClimaNS(Spokoyny et al., 2023)	12.3 (10.6-14.0)	77.5 (75.3-79.5)	84.7 (82.9-86.4)	89.3 (55.6-100.0)	-	-	-	86.00 [†]
ClimaNS (our split)(Spokoyny et al., 2023)	12.8(11.1-14.6)	81.2(79.0-83.2)	77.6(72.9-81.4)	75.8(71.3-80.2)	58.5(55.9-61.0)	45.4(42.7-47.8)	48.4(45.9-51.0)	-
climaQA (Spokoyny et al., 2023)	45.5 (44.9-46.1)	47.7 (47.1-48.3)	44.3 (44.1-44.4)	-	-	-	-	-
climaQA (our split)(Spokoyny et al., 2023)	50.5(49.4-51.6)	50.3(49.2-51.3)	89.5(88.9-90.2)	89.4(88.7-90.0)	78.8(77.9-79.6)	57.4(56.3-58.4)	76.1(75.1-77.0)	-

Table 18: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *a. SVM*

	<i>random</i>	<i>Ours distribution</i>	<i>distilRoBERTa</i>	Spokoyny et al. (2023) <i>ClimateBERT</i>	<i>MiniLM</i>
<i>MRR</i>	0.106 (0.10 - 0.11)	0.343 (0.33 - 0.35)	0.636 (0.63, 0.65)	0.753	0.755

Table 19: Performances on the task of ranking question based on the answer for *ClimaQA*(Spokoyny et al., 2023). *random* ranks randomly each question, *distribution* ranks the question from the most frequent to the least frequent based on the train dataset, *distilRoBERTa* use the outputs of distilRoBERTa for each question/answer pairs to rank the questions. Finally, *ClimateBERT* and *MiniLM* are the performances reported by Spokoyny et al. (2023).

C.9.1 Error Analysis

ClimaINS This dataset contains labels associated to each question, for example “EMISSION” correspond to “Does the company have a plan to assess, reduce or mitigate its emissions in its operations or organizations? If yes, please summarize”. It is balanced, except for “EMISSIONS” being twice as frequent. As shown in Figure 11, the easiest label to identify is the “EMISSION” label, and the hardest are “ASSESS”, “MITIGATE” and “RISK PLAN”.

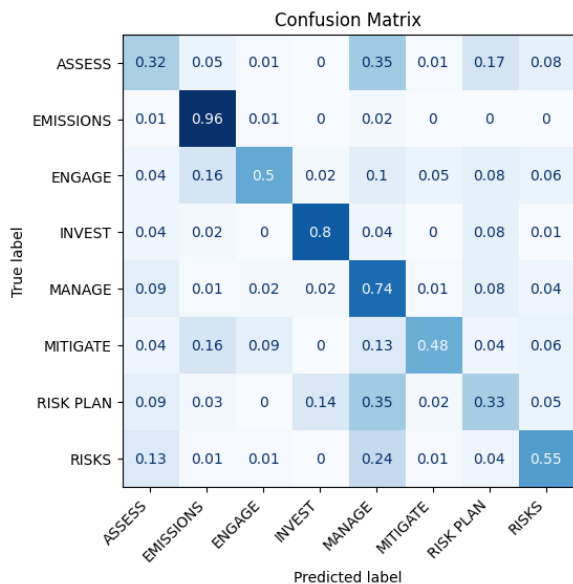


Figure 11: Confusion matrix of GPT-4o-mini predictions on *ClimaINS* dataset.

In the 11 sampled errors, we found 6 instances that are cropped or out-of-context; therefore, they are harder to classify, as they provide fewer context-

ual hints about the original question. Additionally, the average length of text misclassified is shorter than the length of correctly labeled text (respectively 635 and 1091 characters), and more than 54% of text misclassified are shorter than 300 characters compared to only 25% of the correctly classified. This shows that this task is significantly harder with less context and details. We also found one statement referring to its parent company, making it impossible to predict, as well as three statements that are partly or fully off-topic (Example C.75). Overall, we found that most of the sampled errors could answer multiple questions. The dataset is derived from a mandatory survey, which may have led to varying levels of response quality from the companies. Some companies provided uniform responses across all their subsidiaries, while others submitted minimalistic answers, making it difficult to discern which specific questions were addressed.

Example C.75 (MITIGATE: Summarize steps the company has taken to encourage policyholders to reduce the losses caused by climate change-influenced events.). However, Lincoln Financial has adopted electronic document delivery and document delivery suppression initiatives to reduce paper and energy usage.

ClimaQA Instead of using each question as a label, as in *ClimaINS*, in this dataset is a binary classification: given a *response* and a *question*, the model predicts if the response answers the question. Therefore, we reported the F1-scores per question in Figure 12. We see that Q29 to Q38 have F1-scores of either 0% or 1%. This is due to the small size of the subset (less than 5 instances). When fo-

ocusing on larger subsets, we see that most questions have F1-scores above 60%, with the exception of Q13 and Q21. For Q13, the recall is around 25%, and for Q21 the recall is 0%. This is because those questions are about “other targets” which make these questions more open, resulting in varied content. For instance, in Example C.76 the company provided a target about waste management, which is indirectly related to climate-change.

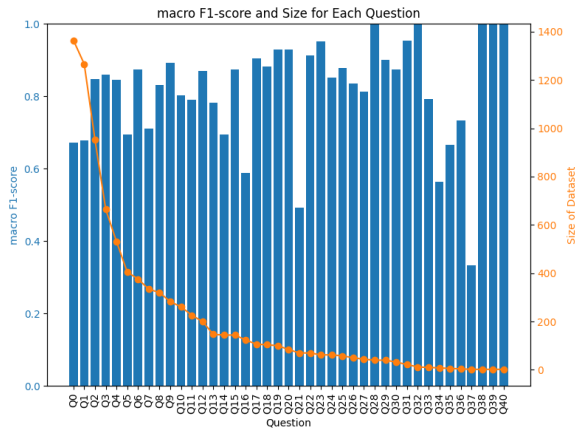


Figure 12: F1-scores (macro) of GPT-4o-mini on *ClimateQA* dataset, grouped by question. **In blue:** The F1-scores. **In orange:** the size of the subset.

70% of FN are questions that actually contain sub-questions. For example, Q3: “Which of the following risk types are considered in your organization’s climate-related risk assessments?” has sub-questions specific to “Market Risk”, “Chronic physical Risk”, etc. When the company answers that a risk is considered, they provide a text specific to that sub-category. The response in Example C.77 was answering the sub-categories: “Acute physical”. However, given the question Q3, one might expect the answer to list all the risks considered, prompting them to classify it as “No (not answering the question)” - as did GPT-4o-mini. For this kind of question, we should construct the dataset differently in order to emphasize that the answer is about a sub-category and often a justification of a value previously reported.

Example C.76. [Yes]

Q13: "Provide details of other key climate-related targets not already reported in question C4.1/a/b."

Answer: "We are targeting a 65% waste diversion rate by 2022 associated in our real estate portfolio."

Example C.77. [Yes]

Q3: "Which of the following risk types are considered in your organization’s climate-related risk

assessments? (Current regulation, Market , Acute physical , Upstream , Emerging regulation, Technology , Legal , Reputation , Downstream , Chronic physical)"

Answer: "[...] At this point, it is difficult to predict and assess the probability of potential risks related to a global warming trend on Dow specifically. Concerns have been raised that climate change may result in sea level rise or average temperature rise. Dow has operations and supply chains in coastal areas that potentially could be affected by these. To mitigate risks associated with severe weather, Dow has engineered the facilities to better withstand climate related events."

C.10 Classification of deceptive techniques:

Task Description The goal is to classify statements into argumentative categories: fallacies, types of arguments, or rhetorical techniques.

Experiment We collected *CC-Contrarian Claims* and *LogicClimate* and reproduced the experiments with the setting described in section B and reported the results in Table 20. Our pre-processing steps did not alter the datasets significantly.

Result Analysis As shown in Table 20, the performances on *LogicClimate* and on *CC-Contrarian Claims* datasets are widely different. The performances on *CC-Contrarian Claims* for fine-tuned models reach F1-scores of up to 71%. On the contrary, for *LogicClimate* the best-performing approaches are the zero-shot models reaching only 27%. In both cases we reached performances in the same order of magnitude as in the original studies, but our approaches underperformed those reported in the original studies.

On *LogicClimate*, our fine-tuned models performed poorly (F1-score between 9-15%), underperforming the random classifier. In the original study, they reported an F1-score of 29.37%. This can be explained by the fact that we only used the domain-specific dataset. Its small size, combined with the label diversity and difficulty, makes this dataset alone less suitable for fine-tuning. Jin et al. (2022) actually fine-tuned the model on general fallacy detection first. For the label difficulty, our TF-IDF baseline performed similarly to the random classifier (13%) suggesting that the labels are not distinguishable through vocabulary only. All this can highlight that the dataset is really challenging, or it could also indicate that there is an issue in

Dataset	Random	TF-IDF	Longformer	DistilRoBERTa	GPT-4o-mini	Llama	Llama 70B	Reference
CC-Contrarian Claims(Coan et al., 2021)	3.2(2.6-4.0)	62.7(60.1-65.1)	71.6(68.5-73.9)	71.2(68.6-73.6)	58.8(55.5-61.3)	40.2(37.0-43.2)	55.8(52.6-58.4)	79 ^j
logicClimate(Jin et al., 2022)	13.3(11.4-15.3)	13.3(8.1-18.0)	15.6(9.6-20.9)	9.4(6.0-12.4)	27.4(15.8-31.4)	10.3(5.4-14.0)	22.1(15.3-28.1)	29.37 ^k

Table 20: Performances of baselines for each dataset using the train, test describe in section 3.3. The performance computed is the macro F1-score. The 95% confidence interval is computed using a 1000 bootstrap of sampled with replace as True of the size of the original test set. The reference performance is the best relevant performances reported in the original papers. *j*. RoBERTa + Logistic Regression, *k*. Electra

the dataset itself. Fallacy detection is highly subjective and IAA agreements are often quite low in that context (Helwe et al., 2024) making them difficult to annotate. Their best-performing models, which can reach 60% on general-domain fallacy detection, only reached 29% on this domain-specific dataset. Our zero-shot approaches reached similar performances (macro F1-score of 27% and micro F1-score of 26% for GPT-4o-mini). The performance gap can also be explained by the source. Indeed, the climate fallacies are taken from real arguments, extracted from a fact-checking website, which are therefore context-dependent and not necessarily explicit, compared to general fallacies which are toy examples extracted from quizzes.

On the contrary, on *CC-Contrarian Claims*, the fine-tuned transformers performed well reaching performances above 70%. This is lower than the performances reported in the original study (79%). This performance gap might be due to the downsizing of the training dataset to 10,000 examples. As there are 18 labels, it resulted in only around 500 examples per label (compared to 1300 originally) and 4 labels appeared less than 200 times in our training dataset. Our TF-IDF baseline reached an F1-score of 62.7%, which is extremely high when considering that there are 18 labels. This shows that the vocabulary used in each argument category is quite specific. In a pilot study, Coan et al. (2021) measured a really low IAA, $\kappa = 0.19$, showing that the task of classifying in a large number of categories might be really difficult for humans. They limited the disagreement by using a decision tree for annotations. They also computed the accuracy of the human annotators on a subset of the dataset for the classification of super-claims. They found an average accuracy above 95% for 4 categories, 86% for the super-claim “Climate movement/science is unreliable”, and only 50% for “No Claim”. This shows that while annotators disagree on the fine-grained categories, they can easily distinguish between broader categories. We also found that most of the errors of GPT-4o-mini occur between sub-claims of the same category (as shown in Figure

13). Both the models and the annotators struggle to classify an argument as “No Claim”. Interestingly, GPT-4o-mini and humans are not biased in the same manner. The model tends to over-classify as “No Claim”, while humans over-classified as “Contrarian Claims”.

C.10.1 Error analysis

CC-Contrarian Claims As shown in figure 13 most of the errors are GPT-4o-mini predicting "No Claim" (42.45%), or errors between the claim types of the same category (31.60%). Only 25.94% of errors are unrelated.

While investigating the errors, we found that 9/30 of statements are multi-faceted, and could fit into multiple categories. Usually, one category is more dominant. Example C.79 could fit in the category of “CO2 is not a pollutant”, but the core argument is “we see no impact of higher CO2 levels on species”. They could also fit into multiple categories because labels are closely linked such as “Weather is cold/snowing” and “Ice/permafrost/snow cover isn’t melting”. In our sample (30 errors), we also found 4 FN which would require more context, such as Example C.78 which could be a contrarian argument by a large emitter advocating against carbon tax; however, it could also be a statement by an activist advocating for stricter regulations. It is not possible to choose without more context. We found 11 texts contain no explicit climate change contrarian claim, however, the arguments can lead to a contrarian claim. For instance, Example C.80 might be part of an unrelated scientific publication, but it could be used in a climate-denial rhetoric to argue that CO2 is actually beneficial.

In 4 cases the model provide the correct reasoning, but decided to use the “No Claim” label anyway. *Overall the labels are really detailed and well define, but due to the granularity the labels are really close together. While we can use the Super-claims labels for evaluation instead of the Sub-Claims, this task could also be a multi-label classification tasks in order to keep the nuanced*

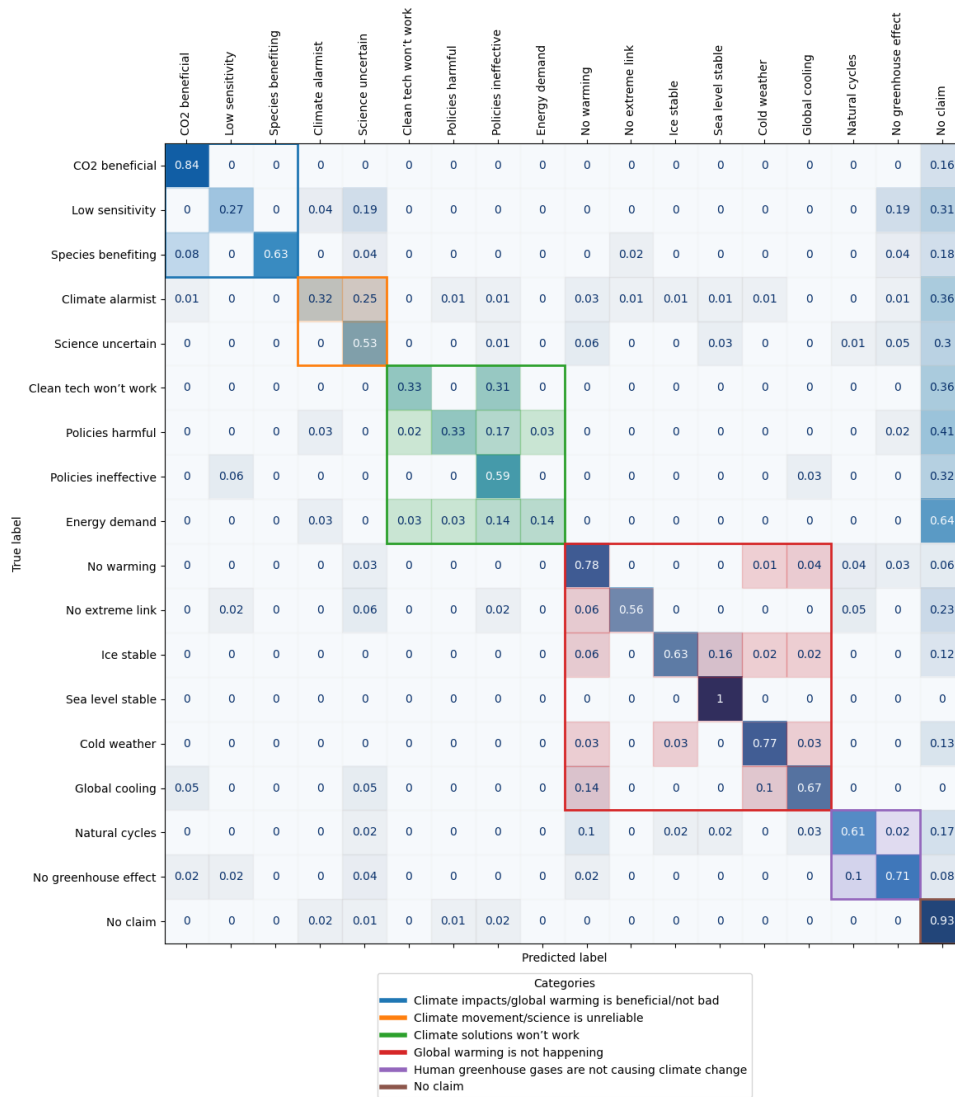


Figure 13: Confusion matrix of GPT-4o-mini predictions on *CC-Contrarian Claims* dataset.

labels.

Example C.78 (No Claim). While a price on GHG emissions would encourage some helpful kinds of innovation, it will not generate other kinds. Coping with climate change will require major breakthroughs in basic science. Such breakthroughs are often elusive, and seeking them is an inherently high-risk venture. The private sector finds it difficult to capture the economic rewards of funding basic science, and placing a price on GHG emissions will not correct this bias. As a result, the private sector usually does not make large, sustained investments in basic science; yet that kind of investment is the key to achieving the breakthroughs needed in climate policy.

Example C.79 (Species/plants/reefs aren't showing climate impacts/are benefiting from climate change). The ocean acidification story depends

only on a chemical hypothesis whereas biological factors can overcome this and create conditions that allow calcification to continue. This is corroborated by the historical record of millions of years of success in much higher CO₂ environments.

Example C.80 (CO₂ is beneficial/not a pollutant). Although elevated CO₂ did not significantly increase total seedling biomass, it did increase it by 14% when averaged across all temperature and fertilization regimes. However, elevated temperature did significantly increase seedling biomass by 55%, when averaged across all CO₂ and fertilization treatments, as did fertilization, by 157%, when averaging across all CO₂ and temperature levels. With respect to root exudation, a similar pattern emerged. Elevated CO₂ did not significantly increase total dissolved organic carbon compounds exuded from seedling roots over a 24-hour period, yet plants

grown in elevated CO₂ exuded 20% more such compounds than ambiently-grown plants did, when averaged across all temperature and fertilization treatments. And once again, elevated temperature and fertilization significantly increased root exudation by 71 and 55%, respectively, when averaged across the other main effect variables.

LogicClimate LOGIC Climate is a dataset that was built to challenge models trained on LOGIC dataset. Therefore, some labels are quite rare, such as circular reasoning (appears once in the test set), equivocation (appears 4 times in the test set) or fallacy of extension (appears 5 times in the test set) which can bias the macro F1-score.

Correctly defining a list of fallacies is not a trivial task, as fallacies have different granularity and can easily overlap (Helwe et al., 2024). Fallacies such as “intentional fallacy” and “Deductive Fallacy” could fit most cases as fallacies are by essence flawed reasoning. While the annotators identified 44 “intentional” fallacies, GPT-4o-mini predicted this fallacy only twice. Moreover, Helwe et al. (2024) concluded that the task of identifying fallacies at a granular level is extremely difficult because of the expertise required and the subjectivity of the task. This explains the low F1-score of both fine-tuned models and zero-shot approaches. Among the errors, we found 3 instances which are not arguments but statements or claims alone. Multiple instances are out-of-context (40%), therefore, missing important information (Example C.81). However, this dataset is particularly interesting due to its construction process relying on fact-checking websites and therefore expert annotations.

Example C.81. Global sea level rose permanently by 1.5 millimeters as a result . [faulty generalization]

C.11 Summary of actual errors

This section contains the details of the actual errors of GPT-4o-mini. These errors are known downfalls of LLMs such as hallucination or being distracted by elements in the prompt (Perković et al., 2024; Shi et al., 2023), but also genuine errors.

Hallucinations and Anchoring We observed a few instances of hallucinations in GPT-4o-mini predictions. It predicted non-existent labels in *SciDCC*, and provided a nonsensical reasoning for a claim-evidence relation prediction in *ClimateFEVER evidence*. Despite observing some, in our

experiments, the model rarely hallucinates. Additionally, GPT-4o-mini is sometimes anchored to a specific word in the prompt (2/9 on *climatext (10K)*) making it misunderstand the instructions.

Redefining terms GPT-4o-mini also struggles when the terms use an alternative definition of known terms (2/16 *climate specificity*, 6/10 in *ClimateEng*, 5/20 on *ClimateBUG Data*). (1) In *ClimateEng*, the label “Politics” is restricted to “leaders, political organizations, policies” therefore excluding activists/protesters. Similarly the label “Ocean/Water” is specific to biodiversity, therefore excluding sea rise. (2) In *climate specificity*, the label “specific” includes all captions and footnotes because they enable readers understand the overall information reported.

Bias In two of the datasets we observed potential biases. On *climate sentiment*, when ambiguity arises (“neutral”), GPT-4o-mini, with our prompt, tends to output preferably the “Opportunity” label (15%) compared to the “Risk” label (12%). The precision for “Opportunity” is 59%, while it is above 80% for both “Neutral” and “Risk”. This is not the case for distilRoBERTa - on “neutral” texts, it predicts “Opportunity” (11%) less than “Risk” (16%). This tends to indicate that the model, with our prompt, has a positive bias. We observed another systematic bias, on *ClimateFEVER*. We observed that GPT-4o-mini, with our prompt, tends to predict disproportionately (1) “NOT ENOUGH INFO” instead of “SUPPORTS”, and (2) “REFUTES” instead of “NOT ENOUGH INFO”. In (1) the claims are all support the existence of climate change; and in (2) the claims all deny climate change. GPT-4o-mini, with our prompt, avoid classifying as “NOT ENOUGH INFO” evidence refuting climate denial claims, but not arguments supporting the seriousness of climate change.

Polysemy On recurrent error is the misunderstanding of the word “sustainable” which is polysemic. It was initially used to mean “consistent” or “viable”, such as in “This business model is sustainable”. However, it is more and more frequently meaning “environmentally and/or socially responsible”. When the term can be interpreted both ways (e.g. “Sustainable business”), GPT-4o-mini, with our prompt, sometimes interprets it as “environmentally and/or socially responsible” (7/17 on *Climate detection*, 5/20 on *ClimateBUG*, 1/8 on *ESGBERT E*). This can be caused by an intrinsic

bias of the model, a more dominant definition in the model’s knowledge, or due to the model being anchored by the prompt which focuses on climate.

Ambiguity We found that the model tends to fail on labels that are more ambiguous, and performs better on labels that are clear-cut binary opposites. (1) 65% of errors in *climate sentiment* occurs on texts with the “neutral” label. (2) The precision of the model on “implicit green claims” is significantly lower than on “explicit green claims” or on “not green” on *Implicit/Explicit Green Claims*. It reached a precision of 55% on “implicit green claims” while all other recall and precision are above 80%. (3) 95% of errors involves “NOT ENOUGH INFO” in *ClimateFEVER evidence*. (4) 70% of errors include “Ambiguous” on *ClimateStance*. (5) 85% of errors involve “neutral” label on *GWSD*. (6) 32% of errors involve labels in the same super-category (compared to 26% different super-category, and 42% for with “no claim”) in *CC-Contrarian Claims*

Context Many of the datasets consist of sentences or paragraphs which often miss context. We found the model struggle with the lack of context as 12.5% of the errors that we sampled were statements out of context or truncated. Moreover, in *ClimaINS*, we found that shorter texts are harder to classify for GPT-4o-mini: misclassified examples are on average shorter than correctly classified ones (635 vs 1091 characters), and 54% of the misclassified examples were shorter than 300 characters, compared to 25% in the correctly classified ones.

D Toolbox

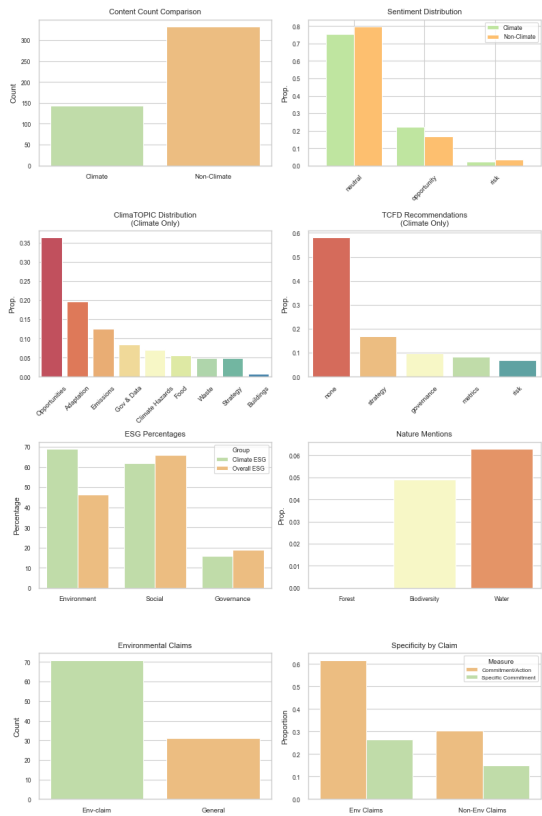
The script is available at https://github.com/tcalamai/acl_climateNLPToolbox. We provide here an example of the toolbox outputs in Figure 14. This was generated with the following simple script:

```
from src.analysis import process_folder,
↳ visualize

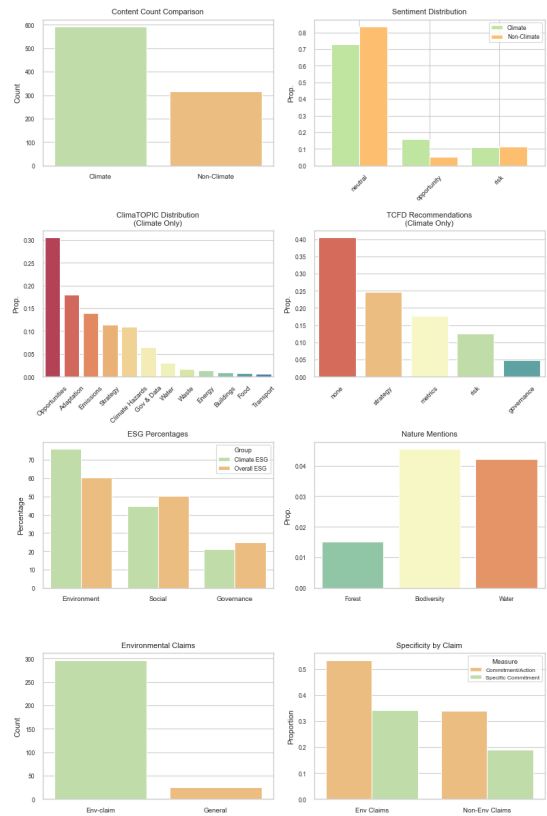
# Path of the folder containing the PDF files
path="case_study/data/"

# Transform PDF into a machine-readable
↳ dataframe:
process_folder(model_name="tfidf", path=path)

# Run all models and generate statistics:
visualize(model_name="tfidf")
```



(a) Company 1



(b) Company 2

Figure 14: Visualization of the statistics generated for 2 companies.