CLIMATEX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements?

Romain Lacombe
Stanford University
rlacombe@stanford.edu

Kerrie Wu
Stanford University
kerriewu@stanford.edu

Eddie Dilworth
Stanford University
edjd@stanford.edu

Abstract

Evaluating the accuracy of outputs generated by Large Language Models (LLMs) is especially important in the climate science and policy domain. We introduce the Expert Confidence in Climate Statements (CLIMATEX) dataset, a novel, curated, expert-labeled dataset consisting of 8094 climate statements collected from the latest Intergovernmental Panel on Climate Change (IPCC) reports, labeled with their associated confidence levels. Using this dataset, we show that recent LLMs can classify human expert confidence in climate-related statements, especially in a few-shot learning setting, but with limited (up to 47%) accuracy. Overall, models exhibit consistent and significant over-confidence on low and medium confidence statements. We highlight implications of our results for climate communication, LLMs evaluation strategies, and the use of LLMs in information retrieval systems.

1 Introduction

The wide deployment of Large Language Models (LLMs) as question-answering tools calls for nuanced evaluation of their outputs across knowledge domains. This is especially important in climate science, where the quality of the information sources shaping public opinion, and ultimately public policy, could determine whether the world succeeds or fails in tackling climate change.

This paper aims to evaluate the reliability of LLM outputs in the climate science and policy domain. We introduce the **Expert Confidence in Climate Statements (CLIMATEX) dataset** [10], a novel, curated, expert-labeled, natural language dataset of 8094 statements sourced from the three most recent Intergovernmental Panel on Climate Change Assessment Reports (IPCC AR6) — along with their associated confidence levels (low, medium, high, or very high) that were assessed by climate scientists based on the quantity and quality of available evidence and agreement among their peers. CLIMATEX is available on HuggingFace and source code for experiments is available on Github.

We use this dataset to evaluate how accurately recent LLMs assess the confidence which human experts associate with climate science statements. Although OpenAI's `GPT-3.5-turbo` and `GPT-4` assess the true confidence level with better-than-random accuracy and higher performance than non-expert humans, even in a zero-shot setting, they, and other models we tested, consistently overstate the certainty level associated with low and medium confidence labeled statements.

With LLMs poised to become increasingly significant sources of public information, the reliability of their outputs in the climate domain is critical for avoiding misinformation and garnering support for effective climate policy. We hope the CLIMATEX dataset provides a valuable tool for benchmarking the trustworthiness of LLM outputs in the climate domain, highlights the need for further work in this area, and aids efforts to develop models that accurately convey climate knowledge.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

2 Prior Work

2.1 LLMs and Uncertainty

Recent literature demonstrates that linguistic statements of certainty impact LLM performance on calibrated NLP tasks. For example, Zhou et al. (2023) [24] find that appending 'weakeners' (i.e. "A: I think it's...") or 'strengtheners' (i.e. "A: I'm certain it's...") to zero-shot QA prompts can significantly impact GPT-3's performance on common QA datasets like TriviaQA. They find that strengtheners surprisingly result in a lower accuracy across their datasets (40%) than weakeners (47%), and suggest that there may be unique challenges with reliably interpreting linguistic cues of high confidence.

To address this issue, and the more general over-confidence problem, recent work attempts to train LLMs to accurately express their own uncertainty. Lin et al. (2022) [12] fine-tune [6] GPT-3 to express how confidently it answers different arithmetic tasks, using both categorical certainty (e.g. "high confidence") and numeric certainty (e.g. "90%"). While they obtain relative success and promising results, the authors note concerns such as over-fitting to their training set distribution.

Kadavath et al. (2022) [8] show that LLMs are capable of 'self-evaluation,' i.e. evaluating their own answers as true or false relative to accepted human knowledge with few-shot learning. They also fine-tune models to predict the probability that they can accurately answer a particular question. The authors train and evaluate many different-sized models using multiple pre-existing generation datasets such as TriviaQA, arithmetic, and code generation. They find that models are initially poor at self-evaluation, but improve with few-shot learning (up to 20 demonstrations).

Prior work has also explored how to evaluate "what LLMs know." Chang et al. (2023) [3] carry what they describe as a "data archaeology" investigation to infer books that LLMs have been trained on, using a "name cloze membership inference query," the task of predicting a masked name in a sentence based on the context surrounding it. Importantly, the human baseline on this task is 0%. The authors find a clear correlation between the frequency at which books appear in datasets over which LLMs are known to have been trained, and their performance on the related cloze task.

Finally, prior work has also included more wholistic evaluations of how LLMs communicate climate-change related information. Bulian et al. (2023) [2] present a framework for evaluating how well LLMs convey climate-change related information. They prompt LLMs to provide a long-form (3-4 paragraphs) response to a climate-related question, and ask humans to rate LLM-generated completions along multiple axes such as accuracy, specificity, completeness, and most relevantly, communicating levels of uncertainty properly.

These prior works suggest that tasking LLMs to assess confidence in climate statements in a zero-shot setting may be a challenging task, particularly for reports released after model knowledge cutoff dates, but it could be made more tractable by few-shot prompting with demonstrations.

2.2 Uncertainty Language in IPCC Special Reports

The IPCC reports have proven to be an excellent context source for retrieval-augmented LLM systems, such as in chatClimate [21], which showed the benefit of using the IPCC reports to ground conversational AI agents. Prior work also used climate science literature as a benchmark for NLP systems, such as CLIMABENCH [11] and CLIMATEXT [22], which evaluate LLMs' performance on a range of tasks like climate topic classifications and knowledge-related QA.

In 2010, the IPCC issued a set of guidelines [14] to lead authors of the IPCC Reports on how to consistently communicate uncertainty. Janzwood [7] analyzes the reports written after the guidelines were published and finds evidence of broad adoption across chapter authors and IPCC reports of the **Confidence** framework, which evaluates scientific confidence in each statement by the quantity and quality of available evidence and agreement among peers. Confidence is measured on a 5-level categorical scale including 'very low', 'low', 'medium', 'high', and 'very high confidence'.

This encouraged us to create the CLIMATEX dataset, comprising of sentences extracted from the IPCC AR6 reports and labeled according to the prevalent 5-levels confidence scale, to evaluate how accurately LLMs assess the confidence level attributed to climate statements by a consensus of human experts. CLIMATEX adds to a growing number of climate benchmarks and is, to our knowledge, **the first LLM climate benchmark to deliberately probe uncertainty, and how accurately these models assess human expert confidence in climate statements.**

3 CLIMATEX: the Expert Confidence in Climate Statements dataset

We introduce **CLIMATEX, the Expert Confidence in Climate Statements dataset** [10], a novel, curated, expert-labeled, natural language dataset of 8094 statements sourced from the three latest IPCC reports (Assessment Report 6: Working Group I [13], Working Group II [17], and Working Group III [18], respectively). Each statement is labeled with its source IPCC report and page number, and the corresponding confidence level on the 5-level confidence scale as assessed by IPCC climate scientists based on available evidence and agreement among their peers.

To construct the dataset, we retrieved the complete raw text from each of the three IPCC report PDFs that are available online using an open-source library [5], normalized the whitespace, tokenized the text into sentences using NLTK [1], and used regular expression search to filter for complete sentences including a parenthetical confidence label at the end of the statement, of the form:

```
``<statement> ({low|medium|high|very high} confidence)''
```

The complete CLIMATEX dataset contains **8094 labeled statements**. From these sentences, we randomly selected **300 statements** to form a **test dataset**, while the remainder form the **train split**. Test samples selection aimed to ensure the test set is: (i) representative of the full data set with regard to confidence class per source report distributions; (ii) representative of the train set with regard to the number of statements from each report; and (iii) contains at least five sentences from each class and each report. Then, we manually reviewed each sentence in the test set, and, where necessary, cleaned up or clarified terms based on paragraph context, to provide for a fairer assessment of model capacity.

We report the percentage of statements per confidence class and source report for the full data set in Figure 3, and for the test set in Table 2. Note that the data set excludes the 'very low confidence' class because 'very low confidence' statements are almost entirely absent from the final reports.

4 Using CLIMATEX to Evaluate LLMs

4.1 Prompting models for confidence in climate science statements

We compared the performance of three recent commercial LLMs: OpenAI's `GPT-3.5-turbo`, `GPT-4`, and Cohere's `Command-XL` [16, 4]. In experiments completed between June 1 and June 9, 2023, we prompted publicly available APIs for each of these models with sentences from the test split of the CLIMATEX dataset, and instructed the model to attempt to predict the corresponding confidence labels using the template shown in Figure 7 in Appendix F.

Using the Demonstrate-Search-Predict (DSPy) library [9], we prompted models in both a **zero-shot setting** with no additional context, and a **few-shot learning setting** with four ground-truth demonstrations randomly selected from the train split of the CLIMATEX dataset (one from each confidence class). Models were prompted to output one of the four confidence levels, or to communicate if they were unable to make a confidence classification due to knowledge limitations.

4.2 Metrics: quantifying confidence levels

To facilitate analysis, we map categorical confidence levels to a numerical score, where 'low' corresponds to 0, 'medium' to 1, 'high' to 2, and 'very high' to 3. The (rare) responses that do not fit into the four confidence level options are excluded from calculations. This mapping allows us to treat our LLM classifiers as regression models, and assess their performance by contrasting predicted confidence levels with ground truth as in a regression setting. We report our results as follows:

- Table 1 compiles the slope and bias of our regression of classifier outputs, for a view of the discernment (how well it distinguishes classes) and over/under-confidence of each model;
- Figure 1 plots the average confidence level predicted by `GPT-3.5-turbo` in the few shot setting for statements from each ground truth class (solid line), breaks down this result for statements from pre-knowledge cutoff date WGI report (dashed line) vs. post-cutoff WGII/III reports (dotted line), and compares it with a perfect classifier (red dashed line);
- Appendix C reports plots for each experiment in Figure 4, and underlying data in Table 3;
- Table 4 in Appendix D compiles classifier performance results for each experiment.

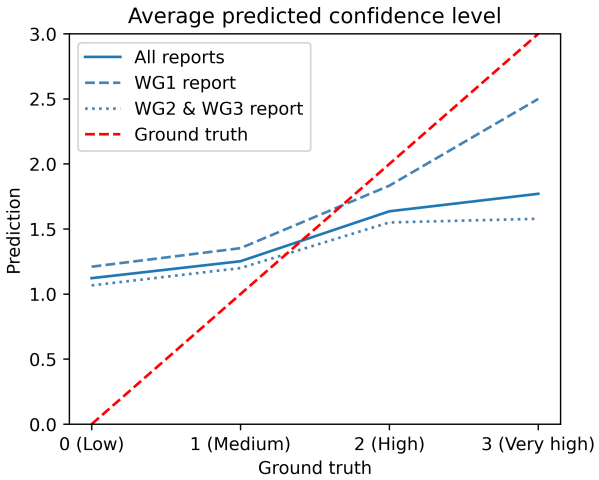

Figure 1: Average predicted confidence level per class. Setting: `GPT3.5-turbo`, few-shot.

Models		**GPT-3.5-turbo**		**GPT-4**		**Cohere Command XL**	
Setting		Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
WGI	Slope	0.368	**0.435**	0.399	**0.445**	0.139	**0.301**
	Bias	+0.198	+0.224	+0.354	+0.255	+0.826	+0.746
WGII/III	Slope	0.184	0.189	0.200	0.297	0.060	0.239
	Bias	-0.136	-0.151	-0.247	+0.023	+0.664	+0.661
Total	Slope	0.215	0.233	0.223	0.323	0.069	0.257
	Bias	-0.046	-0.054	+0.042	+0.083	+0.708	+0.683

Table 1: Regression of average score per category against ground truth.

How to read this table? *(i) Slope: how much the average predicted score varies between statements of adjacent ground truth confidence classes. It measures how discerning the model is. Perfect classifier: 1. Random classifier: 0. (ii) Bias: difference between average predicted confidence and ground truth. It measures any bias towards over-confidence or over-caution. Unbiased classifier: 0. Over-confident by one class (classifies 'medium' as 'high'): +1: Over-cautious by one class (classifies 'medium' as 'low'): –1.*

5 Results and Analysis

We present a summary of our results in Table 1, and an illustration of the confidence classification performance of `GPT3.5-turbo` with few shots demonstrations in Figure 1. We report detailed results for all models in Figure 4 and Table 4 (Appendix D), which lead to the following observations.

(1) Recent LLMs can classify human expert confidence in climate-related statements, but do so with limited accuracy. `GPT-4` can classify statements with up to 44.3% accuracy in the zero-shot learning setting, and 47.0% in the few-shot setting (Table 4). This is slightly better than the 36.3% achieved by a small sample of 3 non-expert humans (Appendix E). All models exhibit a positive correlation between ground truth and predicted confidence level, with a slope varying from 0.06 for `Command-XL` in a zero-shot setup on WGII/WGIII reports (near random classifier), up to 0.445 for `GPT-4` in the few-shot in-context learning setting on the WGI report.

(2) Some models are consistently overconfident. Cohere's `Command-XL`, in contrast to `GPT-3.5-turbo` and `GPT-4`, does particularly poorly by making strongly overconfident classifications (+0.708 bias). Even though `GPT-3.5-turbo` and `GPT-4` perform better overall (biases less than +0.1), they are not perfect—**all three models consistently over-estimate confidence in the 'low' and 'medium' categories within the CLIMATEX dataset (mean score > 1.0)**.

(3) Few-shot learning significantly improves performance for all models. When few-shot demonstrations are sampled from the train set to add context to the prompt, the slopes reported in Table 1 improve for all models, away from 0 (random classifier) and towards 1 (perfect classifier).

(4) Models very rarely convey knowledge limitations. 'Support' in Table 3 reports the number of valid confidence classification, and shows that models convey knowledge limitation ("I don't know") for none to at most 4% of prompts. However, we caution that our prompt begins with, "You are a knowledgeable climate science assistant trained to assess the confidence level associated with various statements," which could influence the models' knowledge-limitation outputs. We conducted robustness checks by prompting non-expert humans with the same task on the CLIMATEX test set, and by prompting LLMs on nonsensical sentences (reported in Appendix E).

(5) Models are significantly better at discerning confidence levels for statements in the WGI report (released before the GPT-3.5-turbo knowledge cutoff date) than in the WGII/WGIII reports (released afterwards). As reported in Table 1, the slope of average prediction scores relative to ground truth is about twice as high on statements from WGI than from WGII/III. Interestingly, this effect is also slightly visible in Cohere AI's `Command-XL`.

Because the training sets for the models we tested are not disclosed, we cannot determine whether this is an emerging capability of LLMs or simply accurate recall of pre-training data. The discrepancy in model performance on the WGI report before the `GPT-3.5-turbo` knowledge cutoff date, compared to the WGII and WGIII reports that were published afterwards, strongly suggests the latter. We cannot exclude either possibility without more transparency on training procedures and datasets.

We also caution that each report covers different topics: the physical scientific basis for climate change (WGI), socio-economic and natural systems vulnerability to climate change (WGII), and climate mitigation options (WGIII). Some performance differences could be explained by relative exposure to literature on each topic during training, rather than recall of the specific IPCC language.

6 Conclusion and Future Work

Our findings add to the body of work calling to accelerate research on improving LLMs' communication of knowledge limitations to users, and calibration of their confidence assessment of statements (from themselves or others). Avoiding 'hallucinations' from 'confidently wrong' LLMs will be key for language model-based applications to function effectively as knowledge retrieval systems. An ability to generate outputs that convey confidence and certainty levels adequately is paramount for LLMs to meet that objective.

This is especially true regarding climate science and the communication of climate knowledge, which shapes public policy and impacts support for climate solutions. In the context of rising political polarization in public policy debates on climate change, where accurate and nuanced climate science communication is crucial, widely-deployed LLMs with the shortcomings highlighted here could exacerbate the spread of misinformation.

Areas for future work to improve and better understand LLM performance on classifying confidence levels associated with climate science statements include: (i) exploring better retrieval methods to select examples or context to include in the prompt; (ii) assessing open-source LLMs on the CLIMATEX test set, before and after fine-tuning them on the train set; (iii) probing whether natural language cues and qualifiers impact model performance on this task; and (iv) conducting a baseline human expert performance assessment on the CLIMATEX dataset.

Acknowledgements

The authors wish to thank Prof. Christopher Potts and Dr. Mina Lee at Stanford University for guidance on this project. We are also grateful to the thousands of climate scientists and experts who collaborated on the IPCC AR6 reports, and the many more whose research informed them. This work and the dataset we constructed would not have been possible without them.

Dataset and code

The CLIMATEX dataset is available for download on HuggingFace.

Code for experiments reproduction and data analysis is available on Github.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[2] Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels Mede, Markus Leippold, and Nadine Strauss. Assessing Large Language Models on climate information, 2023.

[3] Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4, 2023.

[4] Cohere. Cohere's Command Model, 2023.

[5] Mathieu Fenniak, Matthew Stamy, pubpub zz, Martin Thoma, Matthew Peveler, exiledkingcc, and PyPDF2 Contributors. The PyPDF2 library, 2022.

[6] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.

[7] Scott Janzwood. Confident, likely, or both? The implementation of the uncertainty language framework in IPCC special reports. *Climatic Change*, 162(3):1655–1675, October 2020.

[8] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know, 2022.

[9] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP, 2023.

[10] Expert Confidence in Climate Statements (CLIMATEX) dataset. `https://huggingface.co/datasets/rlacombe/ClimateX`, 2023.

[11] Tanmay Laud, Daniel Spokoyny, Tom Corringham, and Taylor Berg-Kirkpatrick. ClimaBench: A Benchmark Dataset For Climate Change Text Understanding in English, 2023.

[12] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words, 2022.

[13] V. Masson-Delmotte, P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[14] Michael D. Mastrandrea, Christopher B. Field, Thomas F. Stocker, Ottmar Edenhofer, Kristie L. Ebi, David J. Frame, Hermann Held, Elmar Kriegler, Katharine J. Mach, Patrick R. Matschoss, Gian-Kasper Plattner, Gary W. Yohe, and Francis W. Zwiers. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties, 2010.

[15] The Onion. The Onion: America's Finest News Source, 2023.

[16] OpenAI. Models, 2023.

[17] H. O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama, editors. *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, USA, 2022.

[18] P. R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, Luz S., and Malley J., editors. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, USA, 2022.

[19] B. Thomas. Onion News Articles Dataset, 2023.

[20] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

[21] Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Nicolas Webersinke, Christrian Huggel, and Markus Leippold. chatClimate: Grounding conversational ai in climate science, 2023.

[22] Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. Climatext: A dataset for climate change topic detection, 2021.

[23] Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation, 2023.

[24] Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models, 2023.

A Appendix: IPCC Guidelines to Authors on Confidence and Uncertainty Communication

In this section we present the IPCC guidelines to authors, as illustrated in the Working Group I Assessment Report 6 [13].

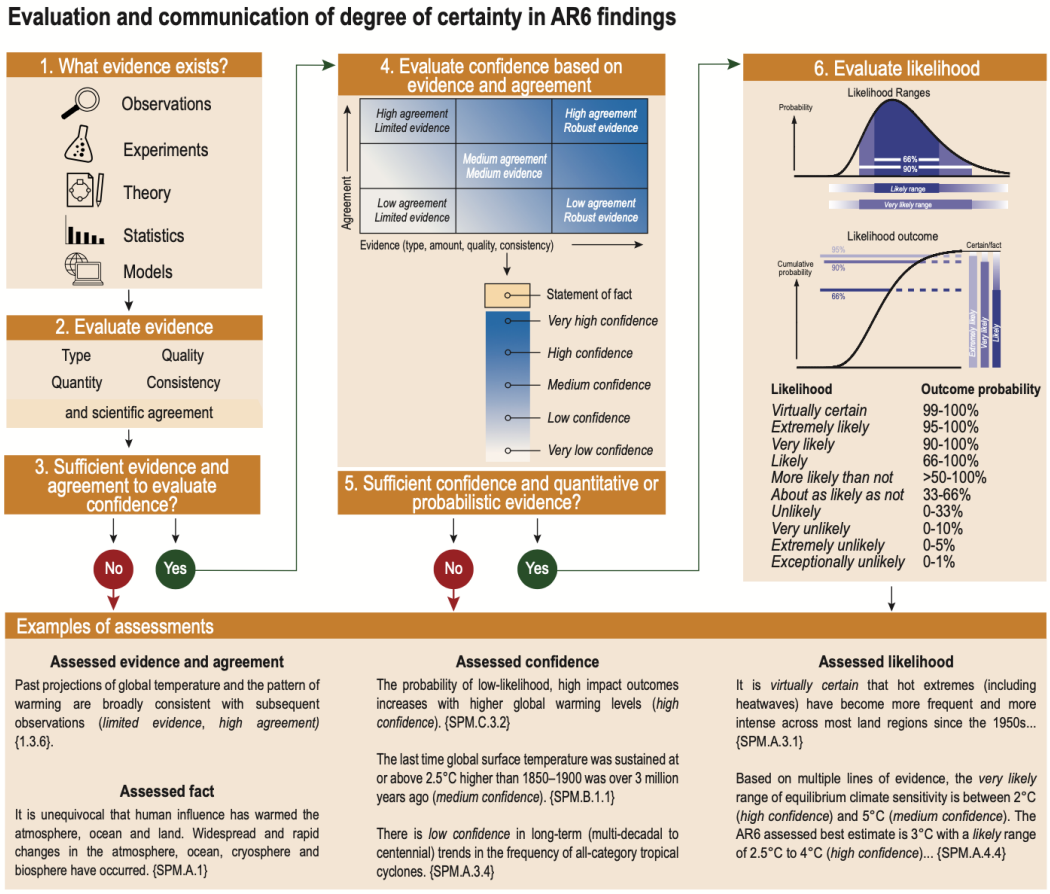

Figure 2: IPCC guidelines to authors on communicating uncertainty and confidence.

B Appendix: CLIMATEX Dataset Confidence Distribution

Report	Low	Medium	High	Very high	Total
WGI	20	35	30	10	95
WGII	25	55	55	35	170
WGIII	5	10	15	5	35
Total	50	100	100	50	300

Table 2: Test set: breakdown by confidence level and source IPCC report.

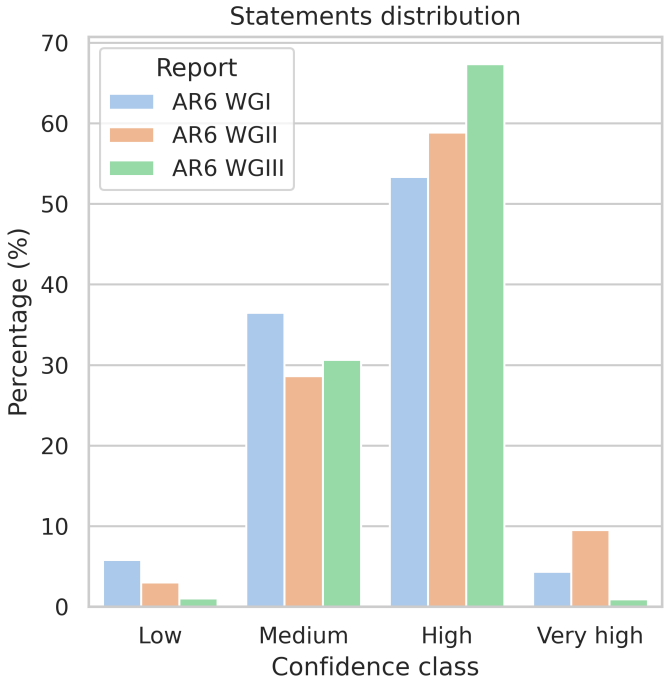

Figure 3: Percentage of statements corresponding to each confidence level and IPCC report source in the CLIMATEX dataset.

C Appendix: Confidence Results Tables and Figures

In this section we present the detailed confidence label predictions results for all experiments.

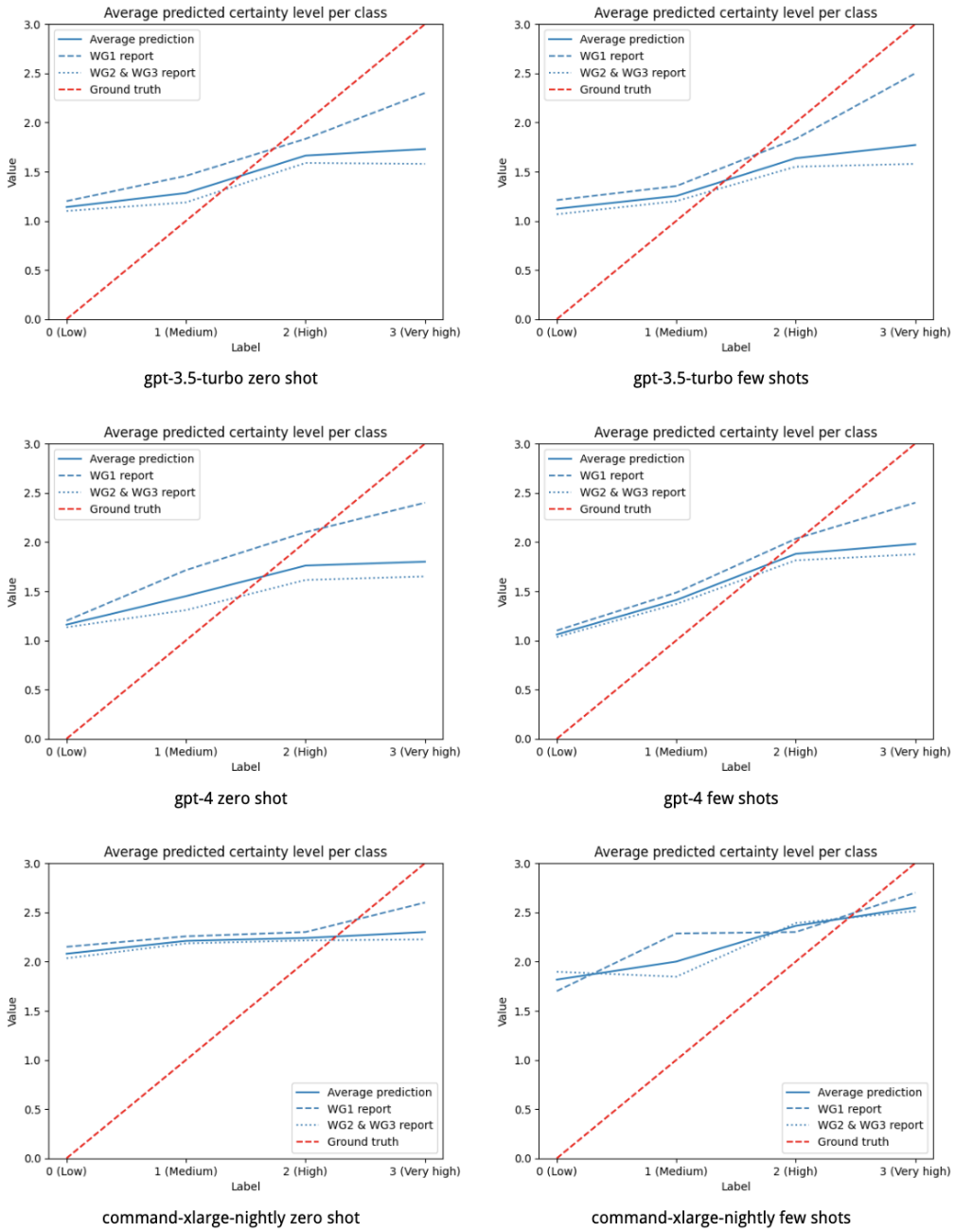

Figure 4: Plots of average confidence level predictions per ground truth class label for all models in all settings.

Models		**GPT-3.5-turbo**		**GPT-4**		**Cohere Command XL**	
Setting		Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
'very high' (3)	WG1	2.30	2.50	2.40	2.40	2.60	2.70
	WG2&3	1.59	1.58	1.65	1.88	2.23	2.51
	All reports	1.73	1.77	1.80	1.98	2.30	2.55
	Support	48	48	50	50	50	49
'high' (2)	WG1	1.83	1.83	2.10	2.03	2.30	2.30
	WG2&3	1.59	1.55	1.61	1.81	2.21	2.39
	All reports	1.66	1.64	1.76	1.88	2.24	2.36
	Support	98	99	100	100	100	99
'medium' (1)	WG1	1.45	1.35	1.71	1.49	2.26	2.29
	WG2&3	1.19	1.20	1.31	1.37	2.18	1.85
	All reports	1.28	1.25	1.45	1.41	2.21	2.00
	Support	99	99	100	100	100	100
'low' (0)	WG1	1.20	1.21	1.20	1.10	2.15	1.70
	WG2&3	1.10	1.07	1.13	1.03	2.03	1.90
	All reports	1.14	1.12	1.16	1.06	2.08	1.82
	Support	50	49	50	50	50	49
Aggregate	Ground truth	1.50	1.50	1.50	1.50	1.50	1.5
	Predicted	1.46	1.44	1.56	1.60	2.21	2.18
	Support (pred)	295	295	300	300	300	297

Table 3: Detailed results: Model average predicted confidence scores for each class within each setting.

D Appendix: Classifier Results Table

Table 4 presents the precision, recall, and F1 score for each classifier, as well as support (number of sentences for which the model answered with a valid confidence label). Note that the 'very high' class and the 'low' class each have 50 total sentences, while the 'high' and 'medium' classes each have 100, for a total of 300 sentences in the test set.

Models		**GPT-3.5-turbo**		**GPT-4**		**Cohere Command XL**	
Setting		Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
'very high'	Precision	0.500	0.476	0.428	0.375	0.221	0.238
	Recall	0.146	0.208	0.120	0.180	0.300	0.592
	F1	0.226	0.290	0.188	0.243	0.254	0.339
	Support	48	48	50	50	50	49
'high'	Precision	0.504	0.485	0.472	0.475	0.332	0.383
	Recall	0.582	0.505	0.680	0.660	0.760	0.546
	F1	0.540	0.495	0.557	0.552	0.462	0.450
	Support	98	99	100	100	100	99
'medium'	Precision	0.389	0.389	0.410	0.466	0.500	0.0
	Recall	0.636	0.616	0.570	0.610	0.010	0.0
	F1	0.483	0.477	0.477	0.528	0.020	0.0
	Support	99	99	100	100	100	100
'low'	Precision	0.167	0.143	0.667	0.833	1.000	0.353
	Recall	0.020	0.041	0.040	0.100	0.020	0.245
	F1	0.036	0.064	0.076	0.179	0.039	0.289
	Support	50	49	50	50	50	49
Aggregate	**Accuracy**	**0.434**	**0.417**	**0.443**	**0.470**	**0.310**	**0.320**
	Macro F1	0.321	0.331	0.324	0.376	0.194	0.270
	Weighted F1	0.384	0.384	0.389	0.430	0.209	0.254
	Support	295	295	300	300	300	297

Table 4: Detailed results: Model classification performance results for the 3 models we assessed in both the zero shot and few shot setting. Reported metrics: accuracy, weighted and macro F1 score, and class-wise recall, precision, and F1 metrics.

E Appendix: Robustness Check

E.1 Baseline dataset

As a robustness check for our approach, we evaluate LLM performance when classifying the confidence of statements outside of climate science and policy using the same experimental setup and analysis methods.

We constructed a small baseline data set consisting of 337 sentences, which includes:

- 100 nonsensical but grammatically correct and complete sentences from the ReCoGS dataset [23]
- 100 confirmed true factoid statements and 100 confirmed false statements from the FEVER dataset [20]
- 37 statements sourced from the Onion News dataset, which may be true, false, or fictional [19, 15]

E.2 Prompting and methods

In addition to using the CLIMATEX dataset, we also prompt the models to classify the confidence levels of statements from our baseline dataset. The purpose of this check is two-fold; (1) check that our experimental method can produce signals differentiating performance between different models and different source materials, and (2) benchmark how models perform on classifying statements unrestricted to the climate science topic from a variety of sources, as a basis for comparing trends that we find when evaluating LLMs on the CLIMATEX dataset.

We use as similar a prompt as possible to the one used for querying the model with CLIMATEX statements, but remove references to climate science and the IPCC reports from the prompt. This minimizes the change in task setup while opening the model and task to a wider range of potential input statements. The complete prompt template is shown in Figure 8.

E.3 Results and Discussion

Although we do not have ground truth confidence labels for our baseline data set, we can use our implicit understanding of the source of each statement to determine appropriate model responses. Nonsensical sentences from the ReCOGS dataset are unverifiable and should elicit a no confidence label or "low" confidence response. Verifiably false sentences from the FEVER dataset should elicit a "low" confidence response (score: 0), while verifiably true sentences should elicit a "high" or "very high" confidence response (score: 2 or 3). Satirical or fictional sentences from the Onion dataset should elicit a "low" or no confidence label response, although some true statements may elicit a "high" or "very high" response.

The results show that:

- Cohere's `Command-XL` is consistently very over-confident on statements from all sources, including verifiably false statements, nonsensical/unverifiable sentences, and likely fictional/satirical statements.
- `GPT-3.5-turbo` and `GPT-4` have relatively high rates (57%, 25%) of refusing to provide a confidence label when prompted with nonsensical ReCOGS sentences; in contrast, `Command-XL` always responds with a confidence label.
- `GPT-3.5-turbo` and `GPT-4` are reasonably able to evaluate the truth of factoid statements from the FEVER dataset, shown by the low mean scores (0.7, 0.3) assigned to false statements and higher mean scores (1.8 and 1.9) assigned to true statements. `GPT-4` is slightly better, assigning lower mean scores to verifiably false statements.

Considering our climate-specific results in the context of baseline robustness check dataset, we see further evidence that some models can discern human confidence levels in specific statements. GPT-3 and `GPT-4` respond to requests for classifying nonsensical sentences from the ReCOGS dataset with no confidence label at a high rate, and can generally differentiate between true statements, false statements, and satirical/fictional statements appropriately. **Consistency in model performance on**

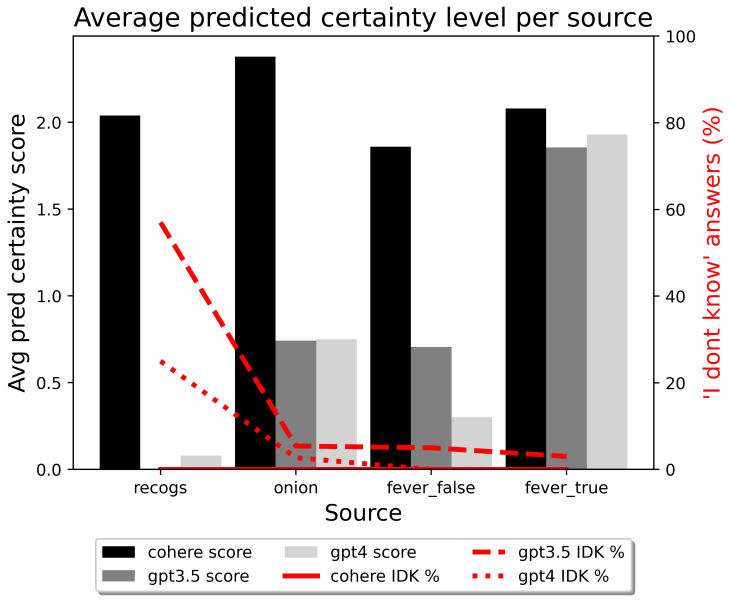

Figure 5: Mean predicted confidence score (left axis) and percentage of no confidence label responses (right axis) from each model, for each source within the robustness check dataset.

both our climate and baseline datasets suggest that our results are likely significant and not random—it is likely that the models' classifications are based on true model "knowledge" from facts and confidence levels seen during training.

E.4 Non-expert human robustness check

Finally, as a baseline study, we performed a cursory non-expert human evaluation. We presented three college-educated (but non-scientist) volunteers with the sentences in the test set and tasked them with classifying the statements according to the 4 confidence classes, and obtained a 36.2% accuracy. Further work is required to evaluate the performance of human experts on this task, and to create a more robust baseline among more volunteers.

	Low	Medium	High	Very high	Accuracy
Non-expert humans	1.46	1.50	1.89	1.88	**36.2%**
GPT-4 (few-shot)	1.06	1.41	1.88	1.98	**47.0%**

Table 5: Non-expert humans vs GPT-4 average confidence level predictions and overall accuracy.

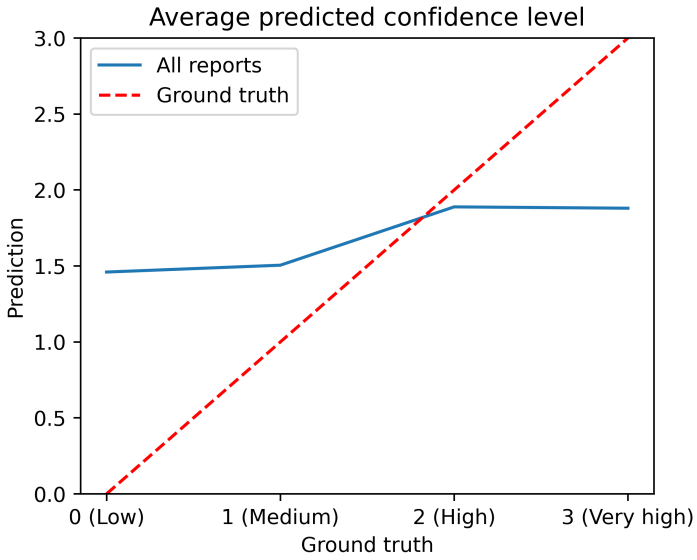

Figure 6: Average confidence level predictions per ground truth class label for the non-expert human evaluation.

F Appendix: Prompts

You are a knowledgeable climate science assistant trained to assess the confidence level associated with various statements about climate change.

You will be presented with a statement about climate science, climate impacts or climate change mitigation which is retrieved or paraphrased from the IPCC AR6 WGI, WGII or WGIII assessment reports. Climate scientists have evaluated that statement as low confidence, medium confidence, high confidence, or very high confidence, based on evidence (type, amount, quantity, consistency) and agreement among their peers. What is their confidence level?

Respond *only* with one of the following words: 'low', 'medium', 'high', 'very high'. If you don't know, you can respond 'I don't know'.

—

Follow the following format.

Statement: ${a short statement about climate.}
Confidence: ${must be *only*: 'low', 'medium', 'high', 'very high'}

—

Statement: Since 1750, increases in CO2 (47%) and CH4 (156%) concentrations far exceed – and increases in N2O (23%) are similar to – the natural multi-millennial changes between glacial and interglacial periods over at least the past 800,000 years
Confidence:

Figure 7: Zero-shot template we used to prompt models for confidence levels associated with climate science statements from the CLIMATEX dataset, along with an example sentence.

You are a knowledgeable assistant trained to assess the confidence level associated with various statements.

You will be presented with a statement. Humans have evaluated that statement as low confidence, medium confidence, high confidence, or very high confidence, based on evidence (type, amount, quantity, consistency) and agreement among their peers. What is their confidence level?

Respond *only* with one of the following words: 'low', 'medium', 'high', 'very high'. If you don't know, you can respond 'I don't know'.

—

Follow the following format.

Statement: ${a short statement.}
Confidence: ${must be *only*: 'low', 'medium', 'high', 'very high'}

—

Figure 8: Zero-shot template we used to prompt models for confidence levels associated with statements from our baseline dataset.