Aligning Unstructured Paris Agreement Climate Plans with Sustainable Development Goals

Daniel Spokoyny*
Carnegie Mellon University
dspokoyn@cs.cmu.edu

Janelle Cai*
MIT
jcai18@mit.edu

Tom Corringham
UC San Diego
tcorringham@ucsd.edu

Taylor Berg-Kirkpatrick
UC San Diego
tberg@ucsd.edu

Abstract

Aligning unstructured climate policy documents according to a particular classification taxonomy with little to no labeled examples is challenging and requires manual effort of climate policy researchers. In this work we examine whether large language models (LLMs) can act as an effective substitute or assist in the annotation process. Utilizing a large set of text spans from Paris Agreement Nationally Determined Contributions (NDCs) linked to United Nations Sustainable Development Goals (SDGs) and targets contained in the Climate Watch dataset from the World Resources Institute in combination with our own annotated data, we validate our approaches and establish a benchmark for model performance evaluation on this task. With our evaluation benchmarking we quantify the effectiveness of using zero-shot or few-shot prompted LLMs to align these documents.

1 Introduction

The 2015 Paris Agreement established 165 country specific Nationally Determined Contributions (NDCs) specifying global commitments to sustainability and resilience. Revised NDCs were released in 2021–2022. The NDCs set ambitious climate action targets but are presented in *unstructured texts* making any analysis or tracking of goals over time difficult. The United Nations Sustainable Development Goals (SDGs) provide a structured framework of 17 goals and 169 sub-targets aimed at promoting global well-being and sustainability. The SDGs serve as a hierarchical taxonomy. Linking NDC text spans to SDG goals and targets can enhance the understanding of global sustainability targets and offers a clear way to track progress. Previous work by Climate Watch at the World Resources Institute manually linked NDC text spans to SDG goals and targets (Northrop 2016) but such an effort is difficult to generalize and maintain as new NDCs are released every five years. This study explores computational methods to tackle the challenge of aligning detailed, jargon-heavy unstructured climate documents to structured taxonomies in the context of limited labeled data, allowing us to significantly extend and enhance existing NDC-SDG datasets.

Prompting LLMs provides a relatively unsophisticated yet powerful way to leverage the models' capabilities. Furthermore, many of today's most advanced LLMs are easily accessible through APIs and web interfaces, making them well suited for a wide range of climate policy researchers. However, there are concerns that LLMs can "hallucinate" and may struggle with understanding context, nuance, and long-term dependencies in text, leading to less coherent or relevant outputs in complex tasks. A formal evaluation of the utility of LLMs for the task at hand is currently lacking.

Our contributions are as follows: 1) We conduct an empirical study of the performance of LLMs and cross-encoder architectures on the task of aligning NDCs to SDGs. 2) We introduce a benchmark for comparing our models, annotators, and the existing Climate Watch dataset. 3) We analyze specific methods to further boost performance on this task.

Finally, we will release the full NDC reports with their predicted SDG alignment as an artifact for the community to use, fostering transparency and ensuring the aims of international agreements are better understood, monitored, and ultimately realized.

2 Related Work

2.1 NDC SDG Linking

Existing research that has explored NDC-SDGs has relied on manual expert annotations. Policymakers across several jurisdictions observe that there is significant overlap between the implementation

*These authors contributed equally to this work.

process for SDGs and NDCs, and that the linking of both policymaking processes increases the efficacy of climate policy design. Northrop et al. (2016) and Brandi et al. (2017) provide detailed evidence for the convergence between SDGs and NDCs. Antwi-Agyei et al. (2018) aim to leverage the alignments and misalignments between West African NDCs and global SDGs to increase the efficacy of West African climate policies.

Due to the effort required to align NDCs with SDGs, most studies are limited in scope: concentrating on a specific geographical region (Antwi-Agyei et al., 2018) or selecting a single or subset of the SDG goals (Gallo et al., 2017; Smith et al., 2023). In our study we have coverage across all SDGs and targets, geographical regions, provided the availability of NDC document in English language, the entire texts of the documents. Other approaches utilize keyword search or extraction techniques to label data, however, these methods have limitations (King et al., 2017), including potential biases introduced by the choice of keywords.

2.2 NLP for Climate

NLP serves as a powerful tool to assist in many climate-related tasks. Stede and Patz (2021) show that NLP can provide many insights to policymakers and activists, as it aids with processing a large quantity with varied types of information. In previous works, NLP has been used to identify climate change misinformation (Farrell, 2019), analyze finance documents for climate-related text (Luccioni et al., 2020), and identify sustainability goals in peer-reviewed academic papers (Smith et al., 2023). Due to the importance of climate documents and the challenges of understanding the technical language used in them, researchers have also trained models specialized for interpreting climate documents, such as ClimateBert (Webersinke et al., 2022) and cross-encoder models to answer questions about climate texts (Spokoyny et al., 2023).

NLP has also been used to align various documents with climate targets, which provide insight regarding progress toward implementation of climate objectives (Roelfsema et al., 2020). Most recently, Juhasz et al. (2024) analyzed climate targets in national laws and policies. They trained a classifier to classify text into three different categories, 'Net Zero', 'Reduction' and 'Other' (Juhasz et al., 2024). Their work demonstrates the potential of using NLP to scale analysis of climate policies. In our study, we extend the climate target classification to the SDGs, which allow us to classify broader targets related to sustainability in climate documents.

There has also been exploratory work on using ChatGPT to interact with climate documents such as the Intergovernmental Panel on Climate Change Report (IPCC) (Vaghefi et al., 2023). In contrast, our aim is to understand how modern LLMs with zero shot prompting, or few-shot in-context-learning could assist in these tasks.

3 Datasets

In this section we will introduce the World Resources Institute's Climate Watch dataset that we used for our experiments and additional benchmarks we constructed (Northrop et al., 2016).

The *Climate-Watch* dataset includes sentences from NDCs submitted before 2021, each of which are labeled with goals and targets. Some sentences are labeled with a single goal and target, and others may also be labeled with multiple goals or targets. Some statistics on the dataset, example sentences and their labels, as well as our data pre-processing can be found in Appendix Section A.1.

3.1 Constructing Additional Benchmarks

We also created two small evaluation datasets that we will use to benchmark various aspects of our prompting strategies.

To construct the *Data-Random* dataset, we preprocess the HTML version of the NDC reports, using the NLTK sentence tokenizer on the the HTML tags that contain the majority of the textual content (<p> and <li>). We further filter the sentences to be between 80 and 300 characters in length. Across all of the reports, this yields over 100,000 sentences. From this set, we randomly sampled 120 sentences to be labeled by our annotators, which yielded sentences from 32 NDC reports.

To construct the *Data-Balanced* dataset, we selected 5 random annotations from *Climate-Watch* for each of the 17 *SDG-Goals*, which represented text from 53 of the NDC reports.

Both of these datasets were subsequently labeled by three separate manual annotators: one expert climate scientist and two university students with some climate policy understanding. Each sentence was independently labeled with up to three *SDG-Goals* that the annotator believed were most relevant to the sentence. For the *Data-Random* dataset,

annotators could optionally select a "not relevant" label if they believed the sentence did not align with any of the *SDG-Goals* .

Annotators were briefed on the 17 SDGs, examined 85 labeled examples from the Climate Watch dataset, and then independently labeled 120 random text spans, associating each with up to three relevant SDGs. We include analysis of Inter Annotator Agreement in Appendix Section A.2.

Later, in Section 4.1 we will use the *Data-Random* to estimate the portion of the NDC documents that have been labeled in the *Climate-Watch* dataset. The *Data-Balanced* dataset will allow us to compare the performance of both our models and annotators against a balanced set of the *Climate-Watch* dataset.

4 Experiments

In this section, we introduce our experiments in which we use different prompting strategies with GPT models to classify sentences according to SDG. We will use ChatGPT-3.5 and GPT-4-Turbo as our main models to conduct prompt-based classification experiments. We will use JSON-mode API option to ensure the model outputs are properly structured for classification tasks. As our zero-shot classification baselines we will use MiniCDP, the cross-encoder model finetuned on the semi-structured Carbon Disclosure Project (CDP) questionnaire data from Spokoyny et al. (2023) as well as its base model architecture MiniLM model.

4.1 *Data-Random*

First, using our manual annotations we will try to estimate the existing coverage of the *Climate-Watch* dataset. We found that out of 120 sentences, 13 were labeled non-relevant by the Expert and 8 were labeled as not-relevant by at least two of the annotators. From this, we estimate that around 85% to 95% of the sentences in the NDC are relevant to some SDG. Since there are on average 724 sentences per document, of which 66.4 sentences are labeled in the *Climate-Watch* dataset, we estimate that only 10-15% of the NDC have been labeled as a result. We show a histogram of the predicted *SDG-Goals* for the *Data-Random* dataset in Appendix Figure 4.

Although this is a very rough estimate, it clearly shows that the vast majority remains unlabeled and motivates the need for a more scalable approach to labeling these documents. Although, to our knowledge, there is no full description of the methodology used to construct the *Climate-Watch* dataset, Northrop et al. (2016) suggests that keyword searches along with possible relevance, such as "countries with large coastlines were initially reviewed to identify alignment with targets relating to oceans and coasts".

Following this analysis, we aim to also measure how LLMs perform compared to our annotators on this random subset of sentences from the NDC documents. To do so we construct a simple prompt to predict multiple *SDG-Goals* for each sentence. We have a simple instruction:

```
Given the following Input Text predict
    the Sustainable Development Goal (
    label) out of the following 17
    options:
```

followed by listing out all of the *SDG-Goals* . We include a full prompt for *SDG-Goal* prediction in Appendix Section C.1. To further encourage the model to produce well-formatted JSON outputs, we include an output specification in the prompt, and we provide multiple numbers for experiments predicting multiple *SDG-Goals* :

```
Generate a json object like so: {\'label
    \': [\'2, 5\']}
```

And lastly, to capture non-relevant sentences, we include "0: None of the above labels are applicable" as an option in the list of *SDG-Goals* as well.

As models we use ChatGPT-3.5 and GPT-4-Turbo with the same prompt. We find that GPT-4-Turbo correctly predicted 6 out of 13 non-relevant sentences, while ChatGPT-3.5 was unable to predict any. Upon closer inspection, we found that ChatGPT-3.5 predicted a very general goal, (Goal-13: Take urgent action to combat climate change and its impacts), for a majority of non-relevant sentences.

To evaluate the performance of the models we calculate the accuracy as whether the model's prediction matched one of the Expert labels. We use the Jaccard similarity to measure the overlap between the sets of *SDG-Goals*. We show the results in Table 1. From the results, we see that on random sentences from the NDC documents, both GPT models perform at similar levels to the annotators. We also conducted an experiment where we prompted the model to only produce one goal. We include these results in Appendix Section B.1.

Annotator	Correct	Wrong	Jaccard
Annotator-1	96	24	0.59
Annotator-2	85	35	0.50
Model	**Correct**	**Wrong**	**Avg**
ChatGPT-3.5	89	31	0.55
GPT-4-Turbo	86	34	0.59

Table 1: Results on multiple *SDG-Goal* prediction for the *Data-Random* dataset.

4.2 *Data-Balanced*

First we want to compare the performance of our annotators against the annotations from the *Climate-Watch* dataset. As our metric, we report whether the percentage of sentences where annotators selected the same *SDG-Goal* as the *Climate-Watch* dataset. For our three annotators we found this to be 49.4%, 57.6%, and 48.2%. By using a balanced dataset, we can also evaluate the average accuracy of our annotators for each *SDG-Goal* shown in Appendix Table 10 along with a confusion matrix in Figure 1.

In Table 2 we compare the performance of our models on the *Data-Balanced* dataset. We find that with the top scoring *SDG-Goal* the MiniCDP model achieves an accuracy of 30.6% while the MiniLM model is almost 9% lower at 21.1%. Both of the LLMs perform much better with the ChatGPT-3.5 model achieving 47.1% and the GPT-4-Turbo model achieving 49.4%.

Annotator	Correct	Wrong	Avg
Expert	42	43	49.4%
Annotator-1	49	36	57.6%
Annotator-2	41	44	48.2%
Model			
MiniLM	18	67	21.1%
MiniCDP	26	59	30.6%
ChatGPT-3.5	40	45	47.1%
GPT-4-Turbo	42	43	49.4%

Table 2: Single *SDG-Goal* prediction results for the *Data-Balanced* dataset.

Since in the *Data-Balanced* split there is only a single *SDG-Goal* label for each sentence, we also aim to quantify how well the models perform against our annotators with multiple *SDG-Goal* label predictions. For the MiniLM and MiniCDP models, we simply take the models' top three scoring goal predictions.

We select the annotator with the highest accuracy against the *Climate-Watch* labels to compare our model predictions against. We use the Jaccard similarity to measure the overlap between the sets of *SDG-Goals* . The results are presented in Table 3.

Annotator	Correct	Wrong	Jaccard
Annotator-1	55	30	0.46
Annotator-2	55	30	0.46
Model			
MiniLM	50	35	0.17
MiniCDP	56	29	0.19
ChatGPT-3.5	58	27	0.48
GPT-4-Turbo	57	28	0.50

Table 3: Multi *SDG-Goals* prediction results for the *Data-Balanced* dataset compared to top performing annotator.

We again find that the MiniCDP model to be slightly better than the MiniLM model with Jaccard scores of 0.19 and 0.17, respectively. While both of the other annotators have Jaccard scores of 0.46, the GPT models achieve higher similarity scores of 0.48 and 0.50.

The confusion matrix in Figure 1 shows high agreement for SDG 13 (Climate Action) but also frequent cross-labeling with other goals, reflecting SDG 13's overarching nature in climate texts. SDG 15 (Life on Land) and SDG 7 (Affordable and Clean Energy) displayed notable confusion with goals concerning water and urban development. In contrast, specific goals like SDG 2 (Zero Hunger) were less represented and often conflated with other poverty and health-related goals. The confusion matrix reflects the SDGs' thematic overlaps, indicating that some noise in annotation is inevitable, even with expert input. Employing LLMs for SDG extraction from climate texts will also entail some acceptable level of noise, consistent with expert-labeled data variability.

4.3 *Climate-Watch*

Although, the *Data-Random* and *Data-Balanced* data splits are relatively small, we have found that prompting GPT models to predict *SDG-Goals* is a promising approach for classifying sentences. In our final set of experiments, we will use the *Climate-Watch* dataset to benchmark prediction of *SDG-Targets*. From the full *Climate-Watch* dataset we randomly selected 200 sentences and in this section will refer to it as the ground truth.

We explore two modes for predicting the *SDG-Targets*, *oracle*: where we use the ground truth

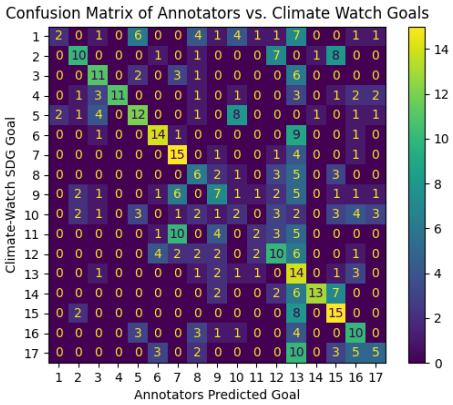

Figure 1: Confusion matrix for the *Data-Balanced* dataset.

SDG-Goal label to sub-select only the corresponding *SDG-Targets*, and *full*: where we predict all *SDG-Targets* for a given sentence. A sample prompt for the *oracle* mode can be found in Appendix Section C.2. We prompt the models to produce the *SDG-Target* labels as JSON objects. Since many sentences have multiple *SDG-Target* labels, for our metric we use the Jaccard similarity. Results for these experiments are shown in Table 4.

Model	Avg	Jaccard
ChatGPT-3.5 *full*	0.385	0.28
GPT-4-Turbo *full*	0.520	0.42
ChatGPT-3.5 *oracle*	0.675	0.49
GPT-4-Turbo *oracle*	0.695	0.57

Table 4: Multi *SDG-Targets* prediction results for the Climate Watch dataset.

For the *full* mode, we see that GPT-4-Turbo is substantially better than ChatGPT-3.5 with Jaccard scores of 0.42 and 0.28, respectively. As expected, in the *oracle* mode both models perform better with the gap between the two models slightly decreasing.

4.3.1 In Context Learning

One of the most desirable features of modern LLMs is their ability to use task-specific examples in their prompt to further boost performance. In the next set of experiments, we additionally provide up to 20 in-context learning (ICL) examples to both of our models. An example of some ICL examples is included in Appendix Section **??**. We show the results in Table 5.

Model	Avg	Number ICL	Jaccard
ChatGPT-3.5	0.40	1	0.31
ChatGPT-3.5	0.46	10	0.35
ChatGPT-3.5	0.49	20	0.36
GPT-4-Turbo	0.56	20	0.44

Table 5: Multi *SDG-Target* prediction results with in-context learning for the *Climate-Watch* dataset.

We find that the ChatGPT-3.5 model improves with additional ICL examples, getting much closer to the performance of the GPT-4-Turbo model. In contrast the 20 ICL examples only slightly improve the performance of the GPT-4-Turbo model.

We also experimented with prompting strategies such as expert prompting but found this did not seem to have any major effect. Results from this experiment are included in the Appendix Section B.2.

4.4 Artifact

To enable climate researchers to continue research in this direction, we use the best existing configuration we identified to annotate the entire NDC documents according to the SDG Goals and Targets. We aim to provide the annotations in a structured format along with the original NDC documents.

5 Conclusion

We have constructed benchmarks to compare the performance of models, annotators, using the *Climate-Watch* dataset on unstructured NDC documents. Using this data we find that existing manual efforts provide low coverage, motivating the need for automated methods. Finally, we found across various experiments that by prompting GPT models we could match the performance of our annotators on *SDG-Goal* and *SDG-Target* prediction. Our findings highlight the potential of leveraging GPT-based models to effectively annotate unstructured climate documents such as the NDCs.

References

Philip Antwi-Agyei, Andrew J. Dougill, Thomas P. Agyekum, and Lindsay C. Stringer. 2018. Alignment between nationally determined contributions and the sustainable development goals for West Africa. *Climate Policy*, 18(10):1296–1312. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14693062.2018.1431199.

Clara Brandi, Adis Dzebo, and Hannah Janetschek. 2017. The case for connecting the implementation of

the paris climate agreement and the 2030 agenda for sustainable development.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Justin Farrell. 2019. The growth of climate change misinformation in u.s. philanthropy: evidence from natural language processing. *Environmental Research Letters*, 14.

Natalya D. Gallo, David G. Victor, and Lisa A. Levin. 2017. Ocean commitments under the Paris Agreement. *Nature Climate Change*, 7(11):833–838. Number: 11 Publisher: Nature Publishing Group.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *ArXiv*, abs/2210.11610.

Matyas Juhasz, Tina Marchand, Roshan Melwani, Kalyan Dutia, Sarah Goodenough, Harrison Pim, and Henry Franks. 2024. Identifying climate targets in national laws and policies using machine learning. *Preprint*, arXiv:2404.02822.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *ArXiv*, abs/2205.11822.

Gary King, Patrick Lam, and Margaret E. Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4):971–988.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing Sustainability Reports Using Natural Language Processing. *arXiv preprint*. ArXiv:2011.08073 [cs].

Eliza Northrop, Hana Biru, Sylvia Lima, Mathilde Bouyé, and Ranping Song. 2016. Examining the Alignment between the Intended Nationally Determined Contributions and Sustainable Development Goals.

Mark Roelfsema, Heleen L. van Soest, Mathijs Harmsen, Detlef P. van Vuuren, Christoph Bertram, Michel den Elzen, Niklas Höhne, Gabriela Iacobuta, Volker Krey, Elmar Kriegler, Gunnar Luderer, Keywan Riahi, Falko Ueckerdt, Jacques Després, Laurent Drouet, Johannes Emmerling, Stefan Frank, Oliver Fricko, Matthew Gidden, Florian Humpenöder, Daniel Huppmann, Shinichiro Fujimori, Kostas Fragkiadakis, Keii Gi, Kimon Keramidas, Alexandre C. Köberle, Lara Aleluia Reis, Pedro Rochedo, Roberto Schaeffer, Ken Oshiro, Zoi Vrontisi, Wenying Chen, Gokul C. Iyer, Jae Edmonds, Maria Kannavou, Kejun Jiang, Ritu Mathur, George Safonov, and Saritha Sudharmma Vishwanathan. 2020. Taking stock of national climate policies to evaluate implementation of the paris agreement. *Nature Communications*, 11(1):2096.

Thomas Bryan Smith, Raffaele Vacca, Luca Mantegazza, and Ilaria Capua. 2023. Discovering new pathways toward integration between health and sustainable development goals with natural language processing and network science. *Globalization and Health*, 19(1):44.

Daniel M. Spokoyny, Tanmay Laud, Thomas W. Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards answering climate questionnaires from unstructured climate reports.

Manfred Stede and Ronny Patz. 2021. The Climate Change Debate and Natural Language Processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Anna Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Nicolas Webersinke, Christrian Huggel, and Markus Leippold. 2023. chatclimate: Grounding conversational ai in climate science. *ArXiv*, abs/2304.05510.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pretrained Language Model for Climate-Related Text. *arXiv preprint*. ArXiv:2110.12010 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Benfeng Xu, An Yang, Junyang Lin, Quang Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *ArXiv*, abs/2305.14688.

A Appendix

A.1 Dataset and Preprocessing

The *Climate-Watch* dataset has the SDG annotations, various associated metadata, and the raw text snippet from the NDC documents. Statistics on the dataset and text snippets are shown in Table 6 and Table 7.

Property	Number
NDC Documents	214
Countries with Documents	186
Labelled Sentences	6813
Sentences with Multiple Goals	1386
Sentences with Multiple Targets	2302

Table 6: Statistics for the Climate Watch dataset.

Each sentence in the document is labeled with one of the 17 SDGs and one of the 169 targets.

Property	Mean
Sentence Length (characters)	137.2
Labelled Sentences per Document	66.4
Goals per Sentence	1.34
Targets per Sentence	1.49

Table 7: Statistics for raw text snippets in the Climate Watch dataset.

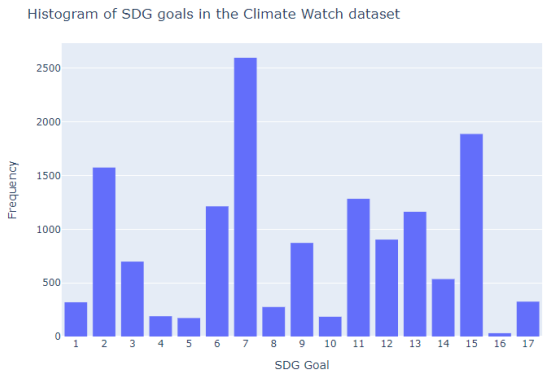

Figure 2: Histogram of the number of labels for each SDG in the Climate Watch dataset.

Some sentences may also be labeled with multiple goals or targets. Example sentences and their labels are shown in Table 8. In Figure 2, we show the distribution of *SDG-Goals* in the *Climate-Watch* dataset.

Additionally, these snippets are not directly linked to the exact locations in the NDC documents. We obtain a dataset of the full texts of the NDC documents as HTML files and using simple heuristics were able to match 94.8% of the annotations to their exact document spans. In Appendix Figure 3 we plot the distribution of where in the NDC documents the *Climate-Watch* annotations are found.

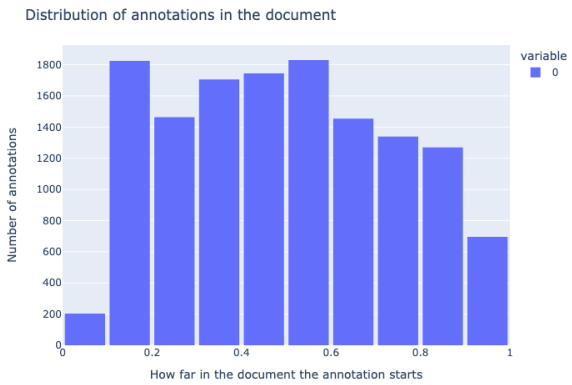

Figure 3: Histogram of where in the NDC documents the *Climate-Watch* annotations are found.

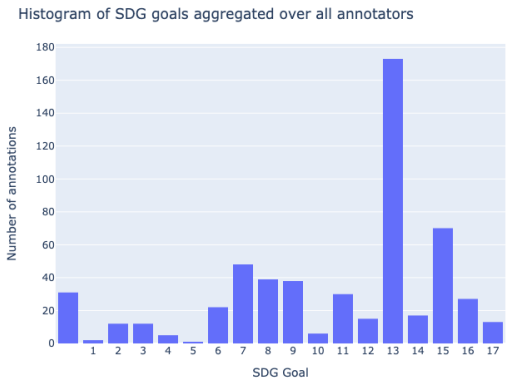

Figure 4: Histogram of the predicted *SDG-Goals* for the *Data-Random* dataset aggregated across all annotators.

A.2 Inter Annotator Agreement

Using our annotators, we show in Appendix Figure 4 the distribution of the predicted *SDG-Goals* for the *Data-Random* dataset, which we can contrast with the distribution of *SDG-Goals* in the *Climate-Watch* dataset (Appendix Figure 3). We found the most common *SDG-Goals* in *Data-Random* were 13, 15, 7 whereas in the *Climate-Watch* dataset it is 7, 15 and 2. The SDG 13 (Take urgent action to combat climate change and its impacts) could be interpreted very broadly and thus our annotators ended up selecting it for a variety of sentences.

SDG 15 focuses on protecting terrestrial ecosystems, SDG 7 targets the provision of sustainable energy, and SDG 2 concerns ending hunger and promoting sustainable agriculture. SDG 13, climate action, is central to the Paris Agreement. SDG 15's frequent appearance stems from the Paris Agreement's emphasis on land use in climate mitigation. SDG 7's prominence aligns with the focus on energy systems transformation in national strategies. The lesser emphasis on SDG 2 in the *Data-Random* dataset compared to the *Climate-Watch* dataset may indicate a different thematic focus in their data set.

For the *Data-Random* split we calculated the inter-annotator agreement using Cohen's kappa (which has a range of -1 to 1) between the expert and each of the novices as (0.629, 0.524) (Cohen, 1960). However, on the *Data-Balanced* the agreement was lower (K = 0.215, K = 0.179), reflecting disparate annotation strategies among the annotators. Notably, some annotators demonstrated a conservative approach, opting to select only the primary goal, whereas others exhibited more leniency in their selections.

Climate Watch Labelled Examples	Goals	Targets
Reduce rural peoples' dependence on fuel for cooking and heating.	12	12.2
Reduce fuel consumption through efficiency standards	7, 11	7.3, 11.2
Guyana will implement other policies to encourage energy efficiency and the use of renewable energy, including building codes and net-metering of residential renewable power.	7	7.2, 7.3

Table 8: *SDG-Goal* and *SDG-Target* labels of example sentences from the *Climate-Watch* dataset.

B Experiments

B.1 *Data-Random* Single Goal Prediction

In this experiment, we prompted ChatGPT-3.5 and GPT-4-Turbo to classify each sentence with a single *SDG-Goal* . We show results in Appendix Table 9. To evaluate the models, we consider a classification correct if it matches any of the *SDG-Goals* that the Expert selected. We find that both models perform well with GPT-4-Turbo being slightly better. We also include the other two annotators as a point of reference although it is not a direct comparison, as annotators were allowed to select up to three *SDG-Goals* .

Annotator	Correct	Wrong	Accuracy
Annotator-1	96	24	80.0%
Annotator-2	85	35	70.8%
Model	**Correct**	**Wrong**	**Avg**
ChatGPT-3.5	87	33	72.5%
GPT-4-Turbo	90	30	75.0%

Table 9: Results on single *SDG-Goal* prediction for the *Data-Random* dataset.

SDG-Goal	Avg	*SDG-Goal*	Avg
7	93.33	8	33.33
15	86.67	9	26.67
6	86.67	17	26.67
13	80.00	1	13.33
5	73.33	11	13.33
3	66.67	10	6.67
2	66.67	12	40.00
4	66.67	14	46.67
16	53.33		

Table 10: Average Annotator Performance by *SDG-Goal* on the *Data-Balanced* dataset.

B.2 Prompting Strategies

There are a variety of prompting techniques that have been shown improve performance such chain of thought (Wei et al., 2022), maieutic prompting (Jung et al., 2022), or self-ask (Huang et al., 2022). Xu et al. (2023) found that providing a model with a prompt that describes an identity of distinguished expert can improve performance. We experiment with a simple form of *expert-prompting* for a climate policy expert. We generated the expert identity using GPT-4 using an example from Xu et al. (2023), and added "You are a climate policy expert..." to the beginning of our instruction. Using the expert prompt, we run *SDG-Target* prediction using both the *full* and *oracle* modes. The results are shown in Table 11 and the full expert-prompt is shown in Appendix Section C.4. We find that there is a small improvement for both models in the *oracle* mode but no effect in the *full* mode.

Model	Avg	Jaccard
ChatGPT-3.5 *full*	0.41	0.27
GPT-4-Turbo *full*	0.51	0.42
ChatGPT-3.5 *oracle*	0.72	0.52
GPT-4-Turbo *oracle*	0.715	0.58

Table 11: Multi *SDG-Target* prediction results with expert prompting on the *Climate-Watch* dataset.

C Prompts

C.1 One Full *SDG-Goal* Prediction Prompt

Below is one full prompt used for zero-shot *SDG-Goal* prediction.

```
Given the following Input Text predict
    the Sustainable Development Goal (
    goal) out of the following 17
    options:
Sustainable Development Goal
1: End poverty in all its forms
    everywhere
2: End hunger, achieve food security and
     improved nutrition and promote
    sustainable agriculture
3: Ensure healthy lives and promote well
    -being for all at all ages
4: Ensure inclusive and equitable
    quality education and promote
    lifelong learning opportunities for
    all
```

```
5: Achieve gender equality and empower
    all women and girls
6: Ensure availability and sustainable
    management of water and sanitation
    for all
7: Ensure access to affordable, reliable
    , sustainable and modern energy for
    all
8: Promote sustained, inclusive and
    sustainable economic growth, full
    and productive employment and decent
     work for all
9: Build resilient infrastructure,
    promote inclusive and sustainable
    industrialization and foster
    innovation
10: Reduce inequality within and among
    countries
11: Make cities and human settlements
    inclusive, safe, resilient and
    sustainable
12: Ensure sustainable consumption and
    production patterns
13: Take urgent action to combat climate
     change and its impacts
14: Conserve and sustainably use the
    oceans, seas and marine resources
    for sustainable development
15: Protect, restore and promote
    sustainable use of terrestrial
    ecosystems, sustainably manage
    forests, combat desertification, and
     halt and reverse land degradation
    and halt biodiversity loss
16: Promote peaceful and inclusive
    societies for sustainable
    development, provide access to
    justice for all and build effective,
     accountable and inclusive
    institutions at all levels
17: Strengthen the means of
    implementation and revitalize the
    global partnership for sustainable
    development

Input Text: Save water for irrigation by
     using plastic films/mulches on
    potato and vegetable fields;
goal:
```

C.2 Sample Hierarchical Prompt

Below is a sample prompt for *SDG-Target* prediction using the *oracle* mode with *SDG-Goal* 7.

```
You are an environmentalist that is
    knowledgeable on the 17 Sustainable
    Development Goals and 169 Targets.
    The following Input Text was
    classified as Sustainable
    Development Goal 7.
Predict the Sustainable Development
    Target (target) out of the following
     options:

Goal 7 Targets:
7.1: By 2030, ensure universal access to
     affordable, reliable and modern
    energy services
7.2: By 2030, increase substantially the
     share of renewable energy in the
    global energy mix by 2030
7.3: By 2030, double the global rate of
    improvement in energy efficiency
7.a: By 2030, enhance international
    cooperation to facilitate access to
    clean energy research and technology
    , including renewable energy, energy
     efficiency and advanced and cleaner
     fossil-fuel technology, and promote
     investment in energy infrastructure
     and clean energy technology
7.b: By 2030, expand infrastructure and
    upgrade technology for supplying
    modern and sustainable energy
    services for all in developing
    countries, in particular least
    developed countries, small island
    developing States, and land-locked
    developing countries, in accordance
    with their respective programmes of
    support
```

C.3 Sample ICL Examples

Below are some sample ICL examples used for *SDG-Target* prediction, for predicting multiple *SDG-Targets*.

```
Input Text: <td><strong>Environmental
    Education and Capacity Building</
    strong></td>
targets: 1.5, 3.d, 4.7, 5.a, 6.b, 7.a,
    8.3, 9.a, 10.3, 11.b, 12.a, 13.3,
    14.a, 15.a, 15.b, 16.b, 17.18
Input Text: Developing and using energy-
    saving construction materials and
    green materials in housing and
    commercial sectors.
targets: 7.b, 9.4
Input Text: Additionally, the Cook
    Islands is looking to embrace proven
     low carbon transport technologies
    and is currently exploring the most
    effective incentives for promotion
    of transition towards clean energy
    transportation.
targets: 7.a, 11.2
Input Text: Increase greenery through
    tree plantation and management of
    gardens and parks.
targets: 11.7
Input Text: 10% of the total population
    (0.8 million beneficiaries (25% are
    women) have increased resilience of
    food and water security, health, and
     well-being in PNG
targets: 2.4, 13.1
```

C.4 Full Expert Prompt

Below is the full expert prompt that we used in our experiments. This was appended to the beginning of each prompt for the expert prompting experiments.

You are a climate policy expert specializing in understanding the complexities of climate systems and the impacts of human activities. Your knowledge spans climate science, mitigation, and adaptation strategies. You excel in analyzing research findings and developing policies that balance scientific evidence, political realities, and societal needs. Your expertise is instrumental in crafting effective and equitable climate policies at all levels, driving action towards a sustainable and resilient future.