

Proportionate Responses: The Case for Small Language Models as Sustainable Expert Climate Policy Annotators

Jetske Bonenkamp, Robbert Biesbroek, Ioannis N. Athanasiadis

Wageningen University and Research, Wageningen, the Netherlands

([jetske.bonenkamp](mailto:jetske.bonenkamp@wur.nl), [robbert.biesbroek](mailto:robbert.biesbroek@wur.nl), [ioannis.athanasiadis](mailto:ioannis.athanasiadis@wur.nl))@wur.nl

Abstract

Tracking progress on climate change adaptation requires systematic and robust annotation of adaptation policies, but manual annotation has become infeasible due to the rapidly growing volume of unstructured documents. Pre-trained language models (LMs) have emerged as promising tools for automating and scaling annotation. However, relatively little is known about their accuracy and quality on complex, free-text annotation tasks. Meanwhile, large LMs (LLMs) are increasingly deployed for large scale text analyses despite their considerable environmental footprint and limited exploration of more sustainable alternatives. By benchmarking model accuracy against human inter-annotator agreement, reviewing perceived annotation quality, and estimating energy and water consumption, we show that fine-tuning small LMs (SLMs) provides a substantially more sustainable alternative to LLMs with only marginal reductions in accuracy and improved perceived annotation quality. Furthermore, we demonstrate the importance of evaluating complex annotation tasks beyond accuracy against human gold standard annotation as ambiguity and human error can obscure meaningful differences in model performance. By reducing the costs of high-quality policy annotation, our approach enables the scalable knowledge extraction needed for tracking progress on adaptation, living evidence syntheses, and evidence-informed policy evaluation.

1 Introduction

Global reporting requirements resulted in a highly voluminous, continuously expanding landscape of climate policy documents. Since the Paris Agreement and its Global Goal on Adaptation (GGA), such documents increasingly cover adaptation *goals* (i.e., statements of government aims and ambitions that define the overall direction of policy) and *instruments* (i.e., tools or resources through which governments seek to achieve their goals).

These goals and instruments, referred to as "policy elements", form crucial evidence for tracking adaptation progress - an issue that has become increasingly urgent in climate debates, particularly following the recently adopted set of 59 indicators to track the targets under the United Arab Emirates Framework for Global Climate Resilience (Rubial et al., 2023; UNFCCC, 2025). These indicators include the adoption and implementation of national adaptation plans, policies, and strategies. The unstructured nature of the evidence (Bonenkamp et al., 2025), combined with the complexity and ambiguity of the concept of adaptation (Orlove, 2022), however, complicates adaptation tracking and has resulted in an incomplete understanding of the current state of adaptation globally. To address this gap, establishing a structured, global-coverage database of adaptation policy at the element level plays a critical role.

Transforming unstructured, multi-sectoral documents into granular, usable evidence traditionally relies on expert text annotation. However, with databases such as Climate Policy Radar already containing over 16,000 climate policy documents published between 2010 and 2025, while the number of documents increases rapidly, manual annotation has become infeasible. Consequently, research has increasingly turned to automated approaches for data processing and analysis (Bonenkamp et al., 2025; Sietsma et al., 2024). Recent advances in the field of Natural Language Processing (NLP), particularly the rise of pre-trained language models, have enabled moving beyond traditional approaches (e.g., text-as-data through topic modeling or text classification) and allow for automating complex text tasks, including detailed annotation processes. In recent years, the use of general-purpose Large Language Model (LLM)s became increasingly common in diverse research settings, given their highly advanced capabilities in many tasks without the need for training data (Wan et al.,

2024). Meanwhile, the high computational demand and, consequently, environmental cost of these large models is often not aligned with their performance gains over solutions with smaller, domain-specific LMs - particularly in big data applications (Wang et al., 2025).

The trade-off between the environmental sustainability of annotation models and the accuracy and quality of automated annotations is difficult to determine in a complex and ambiguous task like annotating adaptation policies. For evaluating the accuracy of NLP models, it is common to rely on hand-labelled "gold standard" data. Expert annotation of free text is inherently complex, granular, and ambiguous, resulting in only moderate inter-annotator agreement (IAA) (Kasner et al., 2025). Consequently, evaluating automated annotation is less straightforward than in, e.g., binary text classification tasks, and a human-labelled gold standard alone may not fully reflect annotation quality.

In this work, we take a critical view on the use of LLMs for large-scale text analysis and assess the potential of Small Language Model (SLM)s as expert climate policy annotators. We consider generative LMs with less than one billion parameters, as well as encoder-only LMs (e.g., BERT models) as SLMs. Although prior work has benchmarked accuracy and computational cost of SLMs for broad NLP tasks with general-purpose (Pham et al., 2025) or fine-tuned (Garg et al., 2025) SLMs, the evaluated NLP tasks are not comparable to the policy annotation task in terms of complexity and domain specificity. As a result, a research gap around the potential of *specialist* SLMs as free-text annotators, particularly in the climate and/or policy domain, has persisted.

To fill this gap, we introduce a framework of fine-tuned, specialist SLMs to extract and categorize adaptation policy elements that have remained hidden thus far. The annotation process is divided into six tasks, each performed by a specific SLM. We benchmark the fine-tuned SLMs against IAA metrics and annotations generated by LLMs across three dimensions: annotation *accuracy*, annotation *quality*, and environmental *sustainability*. In doing so, we assess whether SLMs provide a viable and sustainable approach for upscaling adaptation policy annotation and building a global-coverage database of structured adaptation policy evidence. To assess annotation *accuracy*, we compare model-human and human-human F1 scores while accounting for partial overlap and adaptation intentionality.

The assessment of *quality* uses experts' perceptions and preferences of model- and human-created annotations, determined through a blind review process. Finally, environmental *sustainability* is assessed by estimating the total energy use, carbon footprint, and water consumption that each model would require for processing global-coverage data.

2 Manual annotation

To allow for fine-tuning SLMs, estimate IAA, and measure model accuracy, we created a hand-labelled dataset. First, based on several existing, related typologies, we defined a detailed code book for annotating adaptation policy elements - the full coding scheme is provided in Appendix A. A policy element is based on a text span - a verbatim sequence of words extracted from the original text chunk - and can be assigned type *Goal* or *Instrument*. Within these element types, there are different subclasses (five for *Goal*, six for *Instrument*; see Appendix Table 1). Furthermore, each element can be assigned a hazard and/or sector class (where applicable) out of five classes each, indicating the climatic hazard the element addresses and the sector it targets (see Appendix Table 2). In addition, an adaptation intentionality score is assigned: this 10-point score (0-1) indicates the extent to which the policy element intentionally contributes to increasing resilience (IPCC, 2022) as opposed to indirectly.

Following the code book and building upon a database of pre-classified adaptation-relevant text chunks that average 10 paragraphs (see Bonenkamp et al. (2025) for the details), we extracted a stratified sample of 2,644 chunks used for manual annotation. A team of four trained, graduate-level annotators spent over 600 hours annotating the dataset, including training and iterative refinement of the code book¹. This resulted in identification and categorization of over 12,000 adaptation policy elements. From this hand-labelled dataset, we extract stratified samples: 2,453 (training) and 191 (testing) input chunks and 11,890 (training) and 847 (testing) output policy elements. Among the extracted policy elements, 69% is labelled as *Goal*; 31% as type *Instrument*.

¹The annotation process was conducted by research assistants who were compensated in accordance with standard institutional pay scales and local labor law requirements.

3 Model selection

The annotation framework consists of three specialist models that are all fine-tuned² on the hand-labelled data introduced in [section 2](#). The specialist models ([Table 1](#), column C) complement each other to extract and categorize adaptation policy evidence and are responsible for six different tasks in the process: extracting the policy elements, assigning (sub)types, hazards, and sectors, and predicting adaptation intentionality ([Table 1](#), column A). For each task, the specialist models are benchmarked against human IAA and one or more alternative models ([Table 1](#), column D). For all tasks, we use OpenAI’s *gpt-5-mini* model as a benchmark LLM and prompt it to complete the task while providing the full code book descriptions, but no examples³. For the second task (classifying policy elements as *Goal* or *Instrument*), we additionally benchmark a fine-tuned *ClimateBERT* model to compare the accuracy of the SLM in directly predicting the element type during element extraction.

Considering that the first task (extracting the element text spans) involves language generation, we use a generative SLM for this task. To capture the differences between goals and instruments directly, the model is fine-tuned to return objects containing both the element span and the applicable element type, combining the first two tasks at once ([Table 1](#)). For the first two tasks, we selected the 0.6 billion parameter version of *Qwen 3* ([Yang et al., 2025](#)) as our base model to demonstrate the potential of models with relatively few parameters. This advanced, recent SLM allows full fine-tuning, while required computational resources remain limited. We fine-tune the model to take the full chunks as input and return JSON objects (see [Appendix D](#) for finetuning details) containing the policy element text spans and their element type. We refer to this first SLM as the **annotation model**. For evaluating its accuracy in predicting the right element type, we benchmark the model against the predictions of both *gpt-5-mini* and a fine-tuned *ClimateBERT* model ([Webersinke et al., 2022](#)).

ClimateBERT is a language model that is pre-trained on textual data from the climate domain. In prior, related work, this model has shown performance gains over similar models that are not pre-

trained on climate-related data ([Bonenkamp et al., 2025](#)). As BERT-like models are encoder-only, they are appropriate for tasks that do not involve language generation. Therefore, we use *ClimateBERT* as the base model for our classification and regression tasks and fine-tune two separate models. The first *ClimateBERT* model, the **policy NLI classifier**, covers tasks 3 up to 5. As its name suggests, this model is fine-tuned on data inspired by Natural Language Inference (NLI): it predicts whether the policy element (premise) belongs to the class description (hypothesis). A set of 48,000 premise-hypothesis pairs is used for fine-tuning, comprising sequence-class description pairs with the ground truth class and an *entailment* label (1), as well as sequences paired with an incorrect class, on which the model learns when to predict the *neutral* label (0).

The NLI dataset was composed by adding each policy element in the original training dataset once with its correct subtype label (*entailment*), twice with an incorrect subtype label (*neutral*), with its correct hazard and/or sector where applicable (*entailment*), and a random sample of 5,000 elements that were not assigned a hazard or sector (*neutral*). The class descriptions used for the *neutral* pairs are randomly selected based on the dataset’s class distribution. A few examples of the premise-hypothesis pairs are provided in [Appendix Table 5](#) ([Appendix C](#)). This single-model NLI approach was used to optimize performance on classification of underrepresented classes, given the highly imbalanced training dataset (see [Appendix B](#)). Rather than learning class-specific decision boundaries alone, the model learns to recognize patterns between a policy element and class descriptions, enabling better generalization to classes with relatively few training examples.

For the final task (predicting adaptation intentionality), we fine-tuned another *ClimateBERT* model on a simple regression task, using standardized intentionality scores assigned by the human annotators. This model, the **intentionality predictor**, assigns an adaptation-intentionality score based on the policy element as input. This task can be seen as both a regression, as well as an ordinal classification task: the score can be used to group elements into intentionality classes for policy analyses. For the benchmark approach with *gpt-5-mini*, we therefore map the problem onto three ordered categories (*indirect*, *implicit*, and *intentional* adaptation) and align the LLM-predicted categories to

²All models in the framework are fine-tuned end-to-end. Fine-tuning details are provided in [Appendix D](#).

³A few-shot approach was tested during the initial research phase, showing a substantially increased inference time while accuracy did not improve.

#	A Task name	B Description	C Specialist model	D Benchmark model(s)
1	Element extraction	Extracting policy element text span	Annotation model (SLM 1) Fine-tuned <i>Qwen3-0.6B</i>	gpt-5-mini (JSON mode)
2	Type classification	Assigning element type (<i>Goal</i> or <i>Instrument</i>)	Input: adaptation text chunk ($n=2,453$) Output: list of JSON objects ($n=11,890$)	ClimateBERT, gpt-5-mini
3	Subtype classification	Predicting subtype (e.g., process-oriented goal)	Policy NLI classifier (SLM 2) Fine-tuned <i>ClimateBERT</i>	gpt-5-mini
4	Hazard classification	Identifying hazard(s) addressed, if applicable	Input: pair of premise (element text span) and hypothesis (class description) ($n=48,000$) Output: 0 (neutral) or 1 (entailment)	
5	Sector classification	Identifying targeted sector(s), if applicable		
6	Intentionality prediction	Predicting the adaptation intentionality score	Intentionality predictor (SLM 3) Fine-tuned <i>ClimateBERT</i> (regression) Input: element text span ($n=11,890$) Output: 0-1 intentionality score	gpt-5-mini (ordinal classification)

Table 1: Overview of tasks and models in the annotation process

the scores predicted by the regression SLM. Although this does not allow for a direct accuracy comparison, we can assess the extent to which the two approaches lead to similar categorical interpretations.

4 Evaluation methods

We evaluate the three SLMs (Table 1, column C) across three dimensions: annotation *accuracy*, annotation *quality*, and environmental *sustainability*. In each dimension, we benchmark the models against alternative (L)LMs and IAA metrics that are evaluated through an identical process. Below, we unpack the evaluation process of each dimension.

4.1 Annotation accuracy

To determine the accuracy of the models’ annotations, we take a stratified hold-out sample from the original hand-labelled dataset (see section 2) and use this as the gold standard for evaluating the SLM- and LLM-generated annotations. We compute separate scores to represent accuracy per task (Table 1). Each score is also computed on annotations created by different human annotators: during manual annotation, a set of 287 chunks were annotated by at least two annotators - from which a total of 3,149 policy elements were extracted and coded. As the subset used for multi-coding was determined prior to the manual annotation process, it contains chunks from both the train and test sets. The multi-

coded sample is used to compute benchmark F1 scores of human IAA that the SLMs should ideally approximate. We choose to report F1 scores rather than precision and recall separately, as F1 is symmetric and can therefore be used for both model accuracy and IAA.

For a comprehensive understanding of the accuracy of extracting the text sequences that represent a policy element, we introduce three metrics: *token-level pooled* F1 score (tok_{pool}), *weighted sequence-level pooled* F1 score (seq_{pool}), and *weighted sequence-level macro-averaged* F1 score ($\text{seq}_{\text{macro}}$). To determine whether two policy elements extracted from the same input chunk by different annotators refer to the same entity, we account for partial matches by using a semantic similarity threshold. More details about the metrics and the similarity criteria are provided in Appendix E.

4.2 Annotation quality

Human-labelled data has traditionally been used as the gold standard for evaluating the output of NLP systems. In complex tasks, where IAA scores are typically moderate, such a black-or-white approach can be questionable. To assess how well the annotation accuracy reflects the perceived quality of automatically generated annotations, we conduct a blind review with four experts. Each expert rates a set of 170-250 chunks (ensuring that they do not rate their own annotations), where two variants of the annotated, full chunk were shown. Each vari-

ant was annotated by either a human annotator, the SLM framework, or the benchmark LLM; the rater does not have information on what variant is created by which annotator. Each pair is rated on three metrics: *completeness*, *element type accuracy*, and *element relevance*. Per metric, the raters indicated whether they strongly or mildly preferred annotation A or B, or whether they found them to be of equal quality.

4.3 Environmental sustainability

To estimate the environmental impact of the LM inferences, we use the *EcoLogits* Python library ([EcoLogits](#)). At the time of writing, this library does not account for the impact of input tokens. Therefore, we estimate the weight of the input tokens to be equal to the ratio of costs per input token divided by the costs per output token, based on the *OpenAI* pricing ([OpenAI](#)). This results in an adjusted formula that replaces the number of output tokens used to compute environmental impact:

$$tokens_{total} = 1/6 * tokens_{input} + tokens_{output}$$

Using this adjusted variable gives a total estimation of the differences in impact across different LMs, prompt detail, and model reasoning settings. In addition to the benchmark *gpt-5-mini* model, we estimate the costs of a scenario in which the larger *gpt-5* model would be used to process the entire dataset. To estimate the impact of inference on classification and regression models, as these involve encoder-only models that are run locally, we use the *CodeCarbon* Python library ([CodeCarbon](#)).

5 Results

Here, we discuss the observed strengths and weaknesses of each specialist SLM in the framework separately.

5.1 Annotation model

Element extraction (task 1) - the accuracy scores of both the specialist annotation SLM and the benchmark LLM do not match human IAA in annotating adaptation policy overall, but the differences are relatively minor ([Figure 1](#)). The pooled, token-level agreement between the LLM annotator and the human test set even matches the IAA ([Figure 1a](#)). On all three scores, the benchmark LLM shows to create more accurate annotations than the specialist SLM when taking the human-labelled data as the gold standard. Overall, however, the LLM tends

to over-annotate, as reflected in the number of extracted policy elements: on average, the benchmark LLM extracts more than 7 policy elements per input chunk, while the human annotators extract 4.4. The specialist SLM seems stricter in its inclusion criteria than the LLM, hence an average of 5.5 elements per chunk. As a result, the precision of the specialist SLM is generally higher, but the wide coverage of policy elements by the LLM leads to a substantial advantage on the recall metric and, therefore, a slightly higher F1 score, particularly on the pooled metrics ([Figure 1a,b](#)). The minor difference of seq_{macro} (0.42 vs 0.45) and the relatively large margin to the IAA on that metric ([Figure 1c](#)) emphasizes that both models tend to struggle with lower-density chunks (i.e., texts where only few or no policy elements are found based on the gold standard), which contribute more strongly to seq_{macro} than to the other scores.

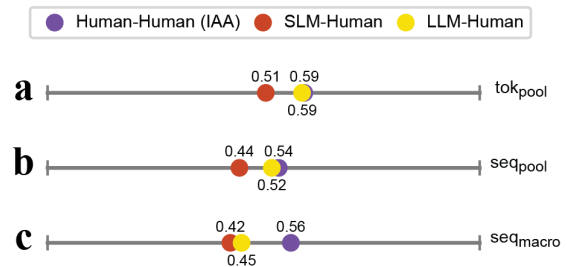


Figure 1: Accuracy and inter-annotator agreement on extracting adaptation policy elements (task 1). The metrics are *token-level pooled* F1 score (tok_{pool}), *weighted sequence-level pooled* F1 score (seq_{pool}), and *weighted sequence-level macro-averaged* F1 score (seq_{macro}). On each metric, the figure shows scores for IAA, the specialist SLM, and the benchmark LLM.

When analysing the *quality* of the generated annotations, the opposite trend emerges: although the benchmark LLM achieves higher accuracy than the specialist SLM, human raters prefer the specialist SLM based on perceived annotation quality. On the *element relevance* criterion, the SLM’s annotations are rated significantly better than those of the LLM; the LLM is rated better on the *completeness* criterion, but this higher average rating is not statistically significant. Although human-created annotations are still rated significantly better than both models, the automatically generated ones are rated as being of equal (SLM - 39%; LLM - 26%) or better (SLM - 17%; LLM - 24%) quality than those of the human annotator ([Figure 2](#)). These substantial numbers underscore the need to step

beyond solely relying on accuracy metrics that take human-labelled annotations as the gold standard.

When it comes to environmental sustainability, a large difference emerges. Per set of 1,000 chunks, the specialist SLM consumes an estimated 0.3 MJ to process the chunks and return JSON objects of the text sequences and their element types (Appendix Figure 1). The gap in energy use to prompting general-purpose LLMs is substantial: when using the smallest variant of OpenAI’s GPT-5 model (*gpt-5-nano*) with low reasoning effort, the consumed energy is estimated to be six-folded (1.6 MJ). Although this still sounds moderate, this model generated poor-quality annotations overall, where the returned text sequences were often not present in the input chunk or not adaptation-relevant. The energy use of the benchmark LLM (*gpt-5-mini* with low reasoning effort) averages 6.8 MJ for processing 1,000 chunks - 25 times as much as the SLM annotator. If one would increase reasoning effort⁴ and/or use the larger *gpt-5* model to, potentially, improve annotation quality or accuracy, energy consumption would increase majorly (Appendix Figure 1). Processing all instances in the pre-filtered, adaptation-relevant dataset (70,000 chunks), consumed energy can approximate 10 GJ as opposed to less than 19 MJ when deploying the SLM annotator (Table 2). Furthermore, extending the prompt with a range of examples would increase input and reasoning tokens, extending the gap even further.

Besides energy use, the differences in water consumption footprint are equally substantial: while the specialist SLM needs approximately 7 liters of water to annotate the full dataset, this goes up to 161 (*gpt-5-mini*, low reasoning) and can even exceed 3,500 liters with the large *gpt-5* model (high reasoning). Figures of the average water consumption, global warming (GW) potential, and (primary) energy use can be found in Appendix F.

Type classification (task 2) - the annotation SLM not only extracted the text sequences: during the same inference, it assigned an element type (i.e., *Goal* or *Instrument*) to each identified element. On this task, the specialist SLM achieves accuracy comparable to the benchmark *ClimateBERT* model (Figure 3a). Both these models exceed the IAA scores, with F1 scores of 0.94 (*Goal*) and 0.83 (*Instrument*) as opposed to IAA F1 scores of 0.92 and 0.80 (Figure 3a). The other benchmark approach,

classification with *gpt-5-mini* (where detailed descriptions of the *Goal* and *Instrument* classes were provided in the prompt) lead to poor results, with F1 scores below 0.6 for both element types (Figure 3a) - particularly due to the LLM being strongly skewed towards the *Instrument* class (Figure 3e). It is important to note that the sample sizes in the confusion matrices (Figure 3b-e) vary as the element type comparisons of the human-human and human-SLM pairs rely on the text spans that were identified by both annotators as extraction and type classification were combined, while the two benchmark models for task 2 (Table 1) took the same set of policy elements as input.

Although these scores may improve when extending the prompt with a varied set of examples for each element type (i.e., few-shot classification) and increasing reasoning, this would further increase energy consumption to disproportionate values for a relatively simple classification task, given that the benchmark *gpt-5-mini* model already has a substantial estimated primary energy use due to the large scale of the element-level dataset (Table 2).

5.2 Policy NLI classifier

The moderate agreement observed across human and computational annotators across the different classes assigned to policy elements (subtypes, hazards, and sectors) highlights the inherent complexity and ambiguity of in-depth adaptation policy annotation. Although the number of matches element pairs in the evaluation set is very low for some of the classes following the high class imbalance (Appendix B), benchmarking the class-level F1 scores of the policy NLI classifier against human IAA and the accuracy of an LLM classifier provides a general understanding of how well the classifier is capable of assigning the correct category to a given policy element.

Among the *Goal* subtypes, the NLI classifier approximates or even exceeds IAA: the NLI classifier and a human annotator mix up similar classes as two human annotators typically do (5a). The robustness of the scores of the *Instrument* subtypes is limited considering the sample size, but the NLI classifier, again, seems to approximate IAA and confuses similar classes as human annotators do (5b,c). Unpacking the most substantial mismatches ($n=94$) between the NLI classifier and the human-labelled standard highlights the capabilities of the model: in the majority (*Goal* - 62%; *Instrument* - 87%) of the evaluated mismatches, review re-

⁴Different reasoning settings for *gpt-5-nano* and *gpt-5-mini* were evaluated on the test set in the initial research phase. No substantial performance gains were observed.

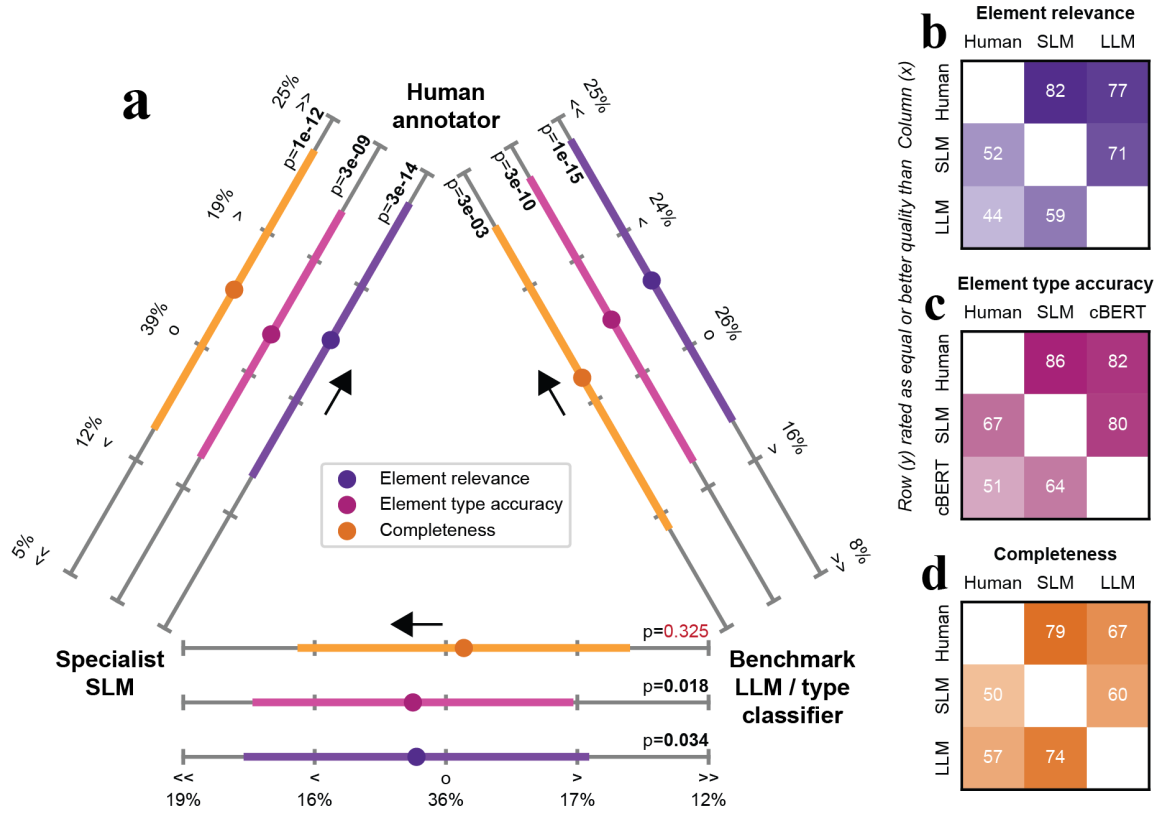


Figure 2: Overview of annotation ratings of a blind, comparative review. The triangle network (a) shows the mean and standard deviation of the 5-point scale ratings. The p-value to indicate whether the preference for a specific annotator (human, SLM, or LLM – indicated by the black arrows) is added to each scale. The percentage of raters selecting each option («, <, o, >, or ») is also indicated. The heatmaps (b, c, d) show the percentage of reviewers rating the row annotator (y axis) to be of equal or better quality compared to the column annotator (x axis), per criterion (element relevance, element type accuracy, and completeness).

vealed that the gold standard was incorrect (see [Appendix H](#) for more details).

On average, human annotators assigned a hazard category to only 14% and a sector category to 35% of the extracted policy elements. The number of instances in the datasets used to determine IAA and accuracy are, therefore, very limited. Among the few matched elements, the human annotators turned out to rarely agree on any explicit hazard being found at all, let alone on the specific category. Although relatively good recall is reported based on the NLI classifier’s predicted hazards and sectors ([Appendix Figure 6a,d](#)), the model links a hazard and sector to a substantially larger share of the policy elements in the test set, particularly to the *environmental* and *meteorological (climate)* hazard classes, resulting in poor class precision ([Appendix Figure 6a,d](#)).

Following the lack of consensus among the annotators, the labelled hazard classes cannot be reliably interpreted as the gold standard. When mapping

(dis)agreement between the NLI classifier and the benchmark *gpt-5-mini* classifier, we observe that the two models agree in most cases, except for the *meteorological (climate)* class, which the benchmark LLM assigns to a larger share of the policy elements to which the NLI classifier did not assign any hazard ([Appendix Figure 6c](#)). Although a reliable comparison between the NLI and LLM classifiers would require additional (blind) review, the excessive estimated environmental cost ([Table 2](#)), given the large number of policy elements in the global-coverage dataset, makes the *gpt-5-mini* model an unsustainable candidate for these classification tasks. Meanwhile, the estimated consumption of the NLI model, despite requiring multiple inferences per policy element to cover all possible classes in premise-hypothesis pairs, remains negligible on the global-coverage dataset ([Table 2](#)).

5.3 Adaptation intentionality predictor

The regression model predicts the adaptation intentionality score with a MSE of 0.02 across the

Task	Model	Energy use (kWh)	Water cons. (L)	GW pot. (kgCO ₂ eq)	Prim. energy use (MJ)
Task 1 (element extraction)	Annotation model (SLM 1)	2	7	<1	19
	Benchmark LLM (gpt-5-mini, low R)	48	162	20	113
	Major LLM (gpt-5, high R)	980	3542	388	9590
Task 2 (type classification)	Annotation model (SLM 1)	N/A (merged in <i>Task 1</i>)			
	Benchmark cBERT classifier	0.05	0	0.01	
	Benchmark LLM (gpt-5-mini, low R)	48	173	25	549
Task 3-5 (sub- type/hazard/ sector classi- fication)	Policy NLI classifier (SLM 2)	0.5	0	0.1	
	Benchmark LLM (gpt-5-mini, low R)	198	712	103	2252
Task 6 (in- tentionality prediction)	Intentionality pre- dictor (SLM 3)	0.04	0	0.01	
	Benchmark LLM (gpt-5-mini, low R)	45	162	23	512

Table 2: Estimated total environmental cost per model. For each SLM and benchmark LLM (reasoning setting indicated by *R*), the table shows the estimated consumption when processing the full dataset of adaptation-relevant chunks (task 1, $n=70,000$) and/or policy elements (tasks 2-6, $n=375,000$)

847 policy elements in the test set, while the score differences among human annotators are larger on average ($MSE=0.039$). Although partial matches in the IAA dataset were taken into account, likely reflecting slight differences in annotator interpretations and, consequently, intentionality scores, the SLM’s substantial advantage over the IAA benchmark shows that it has learned from the underlying patterns in the training data to accurately predict adaptation intentionality. This capability is invaluable to determine whether policy addresses adaptation intentionally, implicitly, or indirectly.

When prompting *gpt-5-mini* to perform ordinal classification and mapping the classes onto the human-assigned (standardized) intentionality scores, we see that the LLM classifies most elements as *intentional* - even those that the human annotator assigned a relatively low score (Appendix Figure 7a). The predictions of the SLM and *gpt-5-mini* are more aligned, particularly showing a clearer distinction between indirect and implicit adaptation on the SLM’s intentionality scale (Appendix Figure 7b). Again, the difference in environmental cost between the specialist SLM and benchmark LLM are large (Table 2) - predicting the

intentionality scores with the *ClimateBERT* regression model leads to very minimal energy (<0.01 kWh) and no water consumption.

6 Discussion

Now that even smaller LMs have gotten so advanced, relying solely on accuracy metrics based on human-created gold standards has become questionable. Our results reveal how challenging achieving high IAA is on a complex, free-text annotation tasks - despite extensive training and iterative fine-tuning of the code book. Considering this moderate overall agreement, a computational annotator that perfectly matches the human gold standard, comprising annotations from multiple annotators, is not a feasible target here. Nonetheless, we show that LMs reach annotation accuracy that approximates IAA and produce annotations that are most often equal to - or even exceed - human-created annotations on perceived quality. Although overall accuracy scores of using a large, general purpose LLM for extracting policy element candidates come close to human IAA compared to a fine-tuned SLM, the consistency and stricter inclusion criteria

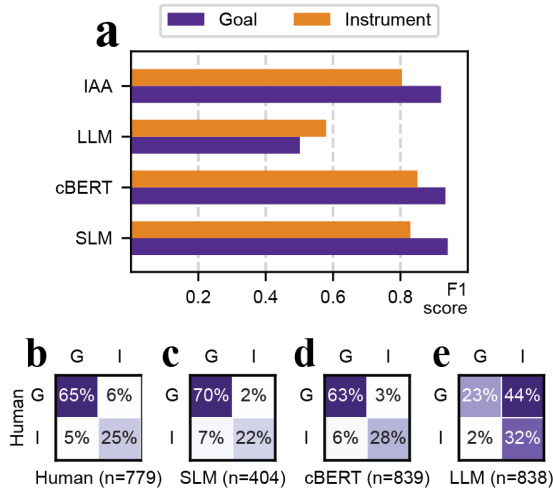


Figure 3: Accuracy and IAA on predicting element types (task 2). The bar chart (a) shows the F1 score of each element type (Goal and Instrument). The confusion matrices (b-e) show the percentages of agreement and confusion across human annotators (a) and between a model and human annotator (c-e).

of the specialist model makes its annotations being perceived as better quality overall. These observations underscore the importance of considering broad evaluation criteria that go beyond traditional accuracy metrics.

In addition to its perceived quality advantage, using a specialist SLM for this relatively complex annotation task brings a substantial advantage in terms of environmental cost: according to the *EcoLogits* estimations, processing all 70,000 chunks from a global-coverage evidence base with the specialist annotation model (SLM 1) as opposed to using *gpt-5-mini* saves 43 kWh - equivalent to approximately one week of electricity use by an average Dutch household (Centraal Bureau voor de Statistiek, 2025) - and more than 150 liters of water. Based on our accuracy and quality evaluations, the relatively small *gpt-5-mini* is not sufficiently advantageous over the specialist SLM to justify this cost. Opting for the larger and more advanced *gpt-5* model, however, leads to excessive and unsustainable consumption that could power a multitude of Dutch households for a full month and cover several days of water use.

Environmental cost reaches even higher levels when considering the use of LLMs for detailed classification tasks. Even when using a relatively small general-purpose LLM like *gpt-5-mini*, energy and water use and global warming potential becomes extremely disproportionate. All fine-tuning pro-

cesses add only 1 kWh energy use combined (see Appendix D), further emphasizing the major gap in environmental cost between specialist SLMs and deploying general-purpose LLMs. This emphasizes that researchers should be aware and critical of the implications of using excessively large LLMs in their research - particularly for structured tasks and in domains where evidence is even more voluminous (e.g., climate change mitigation). Frequent updating of the evidence base, which is inevitable given the continuous development of new climate policy documents and a field that moves towards living evidence platforms and databases that are updated frequently and automatically (Ford et al., 2025), would mean that environmental cost expands even further with every update cycle.

Taking into account the core intention of large-scale annotation of adaptation policies, where addressing climate change and sustainability logically plays a major role, we conclude that bigger is not better in the context of climate (adaptation) policy annotation. Considering that our specialist annotation model is only marginally behind human annotators in terms of accuracy and quality, we confidently suggest the use of fine-tuned SLMs as a sustainable, consistent, transparent, and accurate opportunity for upscaling climate policy annotation.

For related applications, it is important to note that fine-tuning an SLM to become an expert annotator requires sufficient hand-labelled examples. Establishing such a dataset may not always be feasible. In such cases, an alternative approach could be to use an advanced LLM to create a set of training examples - of which a representative sample should be manually checked - and use this to fine-tune an SLM. Depending on the size of the full dataset that needs to be annotated, this approach can majorly decrease environmental and financial costs.

Leveraging our proposed framework of SLMs as expert policy annotators moves beyond the constraints of manual expert annotation and enables processing adaptation policy data at large scale. Through this, it is possible to establish a global-coverage, structured evidence base that allows for systematically tracking adaptation policy progress across the globe, invaluable for global assessment such as the UNFCCC global stocktake (2028) and IPCC AR7 (2028-2029).

Limitations

There are some limitations to our study that are important to consider. As we relied on matched policy elements (i.e., text sequences identified by two annotators) to compute subtype, hazard, and sector IAA, while the classes were imbalanced (see [Appendix B](#)), the data used for testing the NLI model was limited. For a more robust estimation of IAA and model accuracy, particularly on the hazard classes, it is recommended to further expand the test dataset by double-coding a pre-defined set of policy elements.

As the parameters of the *OpenAI* LLMs are not publicly available, the environmental cost of inference with these models was based on estimations. The actual environmental cost may, therefore, somewhat deviate from the numbers presented in our paper. Selecting a different LLM that is not a GPT-5 model, may also influence the results. Furthermore, we have used a rough estimation of the cost of fine-tuning the SLMs in our comparison, as this was not measured during the processes. However, even with a large deviation, the environmental cost is still negligible considering the evident, disproportionate consumption of LLMs.

Finally, we have not accounted for uncertainty in the generated annotations. Uncertainty quantification during manual annotation and in the computationally generated annotations (e.g., by specifying confidence scores), however, could bring an interesting perspective on certain cross-annotator (mis)alignment patterns.

Acknowledgements

JB and RB acknowledge funding support from the Dutch Research Council (NWO VI.Vidi.211.132). Furthermore, we express our gratitude to the team of annotators - Robin Blersch, Denatalie Chrisdamera, Demi Ursem, and Suze Zander - and thank them for their important contributions to establishing the hand-labelled dataset. Lastly, we thank the anonymous reviewers for their detailed review on a previous version of this manuscript.

References

Jetske Bonenkamp, Robbert Biesbroek, and Ioannis N. Athanasiadis. 2025. Transforming adaptation tracking: Benchmarking Transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text. In *Proceedings of The 2nd*

Workshop of Natural Language Processing meets Climate Change.

Centraal Bureau voor de Statistiek. 2025. [Energieverbruik particuliere woningen; woningtype en regio's](#). Last Modified: 2025-09-17T02:00:00+02:00.

Climate Policy Radar. [Climate Policy Radar | AI for climate law and policy research](#).

CodeCarbon. [CodeCarbon](#).

EcoLogits. [EcoLogits | Estimate and track the environmental footprint of GenAI models at inference](#).

James D. Ford, Robbert Biesbroek, Lea Berrang Ford, Felix Creutzig, Neal Haddaway, Sherilee Harper, Jan C. Minx, Mark New, Anne J. Sietsma, Carol Zavaleta-Cortijo, and Max Callaghan. 2025. [Recommendations for producing knowledge syntheses to inform climate change assessments](#). *Nature Climate Change*, 15(7):698–708.

Anandita Garg, Uma Gaba, Deepan Muthirayan, and Anish Roy Chowdhury. 2025. [Emissions and Performance Trade-off Between Small and Large Language Models](#). In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 515–520.

Daniel Henstra. 2016. [The tools of climate adaptation policy: analysing instruments and instrument selection](#). *Climate Policy*, 16(4):496–521. [_eprint: https://doi.org/10.1080/14693062.2015.1015946](#).

IPCC. 2022. *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1 edition. Cambridge University Press.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondřej Plátek, Dimitra Gkatzia, Saad Mahamood, Ondřej Dušek, and Simone Balloccu. 2025. [Large Language Models as Span Annotators](#). *arXiv preprint. ArXiv:2504.08697 [cs]*.

OpenAI. [API Pricing](#).

Ben Orlove. 2022. [The Concept of Adaptation](#). *Annual Review of Environment and Resources*, 47(1):535–581.

Ana Petek, Nikola Baketa, Anka Kekez, Marko Kovačić, Mario Munta, Krešimir Petković, Marjeta Šinko, and Borna Zgurić. 2021. [Unboxing the vague notion of policy goals: Comparison of Croatian public policies](#). *European Policy Analysis*, 7(2):451–469.

Nghiem Thanh Pham, Tung Kieu, Duc-Manh Nguyen, Son Ha Xuan, Nghia Duong-Trung, and Danh Le-Phuoc. 2025. SLM-Bench: A Comprehensive Benchmark of Small Language Models on Environmental Impacts. *Findings of the Association for Computational Linguistics: EMNLP 2025*.

- Pilar Bueno Rubial, Anand Patwardhan, María Luz Falivene Fernández, Joel González, Victoria Laguzzi, Susana Zazzarini, and Carolina Passet. 2023. UAE Framework for Global Climate Resilience: key outcomes of COP28 and pathways towards COP29 and COP30.
- Isabel Segura-Bedmar, Paloma Martinez, and Maria Herrero-Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Seventh International Workshop on Semantic Evaluation*, volume 2, Georgia.
- Sebastian Sewerin, Lynn H. Kaack, Joel Küttel, Fride Sigurdsson, Onerva Martikainen, Alisha Esshaki, and Fabian Hafner. 2023. Towards understanding policy design through text-as-data approaches: The policy design annotations (POLIANNA) dataset. *Scientific Data*, 10(1):896.
- Anne J. Sietsma, James D. Ford, and Jan C. Minx. 2024. The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, 14(1):31–39.
- UNDRR. 2020. Hazard definition and classification review: Technical report. Technical report.
- UNFCCC. 2025. Final list of potential indicators, UAE-Belém work programme on indicators.
- Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 2024. TnT-LLM: Text Mining at Scale with Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5836–5847, New York, NY, USA. Association for Computing Machinery.
- Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuha Lu, Wanjin Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025. A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):145:1–145:87.
- Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pre-trained Language Model for Climate-Related Text. *arXiv preprint*. ArXiv:2110.12010 [cs].
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs.CL].

Appendices

A Class descriptions

Subtypes

Each policy element can be of type *Goal* or *Instrument*. Within these element types, subtypes are defined, which are described in detail in Appendix Table 1. The *Goal* subtypes are based on the policy categories that Petek et al. (2021) defined. The *Instrument* subtypes combine elements of the resources of Henstra (2016) and the taxonomy of POLIANNA (Sewerin et al., 2023).

Hazards and sectors

Policy elements can be assigned a climatic hazard or sector, where applicable. These hazard and sector classes are provided in Appendix Table 2. The typologies are based on the general hazard clusters of the UNDRR (2020) and the impacts of climate change on ecosystems and human systems from IPCC’s Sixth Assessment Report (IPCC, 2022).

B Class distribution

An overview of the distribution of classes is provided in Table 3 (subtypes) and Table 4.

Goal (69%)		Instrument (31%)	
INS	3138 (39%)	ORG	2436 (68%)
PRO	2831 (35%)	KNO	598 (17%)
SEC	1987 (24%)	TEC	213 (6%)
VAL	107 (1%)	REG	153 (4%)
EVA	78 (1%)	PUB	103 (3%)
		FIN	100 (3%)

Appendix Table 3: Distribution of subtype classes in the hand-labelled dataset

Hazard		Sector	
None	10280 (86%)	None	7755 (65%)
MET_c	747 (6%)	NAT	1061 (9%)
GEO	411 (3%)	FOO	992 (8%)
BIO	167 (1%)	WAT	876 (7%)
MET_w	161 (1%)	BUI	764 (6%)
ENV	124 (1%)	HEA	442 (4%)

Appendix Table 4: Distribution of hazard and sector classes in the hand-labelled dataset

Type	Subcategory	Definition
Goal		<i>Quantitative or qualitative objective that a party aims to achieve, with the overarching aim of adapting to climate change. Involves a desired/aspired outcome or intended state/action</i>
	SEC: Sector-oriented	Goals directly aimed at achieving a desired state and/or solving a problem in a specific area of interest, such as a (social) group or region (accessibility and security of goods and services; well-being and prevention; development and mobility)
	PRO: Process-oriented	Goals that focus on the organizational aspect of the regarding policy, only indirectly intended to address an issue (cooperation; capacity-building; strategic planning)
	EVA: Evaluation-oriented	Goals that involve meeting a set evaluation criterion in a given sector (quality; sustainability; effectiveness)
	INS: Instrument-oriented	Goals that refer to the introduction, change, or continuation of a policy instrument as a policy goal (regulation; investment; (re)organization; knowledge and awareness)
	VAL: Value-oriented	Goals that aim at ensuring or advancing core constitutional values of democracies or political fundamentals (justice and law; environmental protection; rights and equality)
Instrument		<i>The means to achieve the defined targets, typically in the form of a tool, mechanism, or strategy. Specifies how a goal can be achieved and can be written in present or past tense</i>
	REG: Regulation and legislation	The use of the state's legal authority to order, forbid, or permit certain behaviour
	FIN: Financial incentive	Taxes, subsidies, or tradable allowances to stimulate behaviour or activities
	ORG: Organization and cooperation	The use of governmental resources through programmes, services, plans, and/or (voluntary) agreements
	PUB: Public investment	Strategic investments and the management thereof, done by public bodies
	KNO: Knowledge and information	Research, education, training, and knowledge sharing, focused on knowledge generation and increasing awareness
	TEC: Technology and innovation	Creation, advancement, and commercialization of new technologies and innovative systems or incentives

Appendix Table 1: Overview of the subtype typology for annotating adaptation policy.

Type	Subcategory	Definition
Hazard		<i>A policy element gets assigned a hazard only if one can clearly identify what type of hazard it addresses by the text sequence itself or the context (i.e., the surrounding text in the same chunk)</i>
	MET_w: Meteorological (weather)	Addresses problems rising from extreme weather events produced by the Earth's state and behaviour, occurring during a short period (minutes to days) and/or on small scale (storms, extreme temperatures, heavy rain/snowfall, heatwaves, short-term droughts)
	MET_c: Meteorological (climate)	Addresses problems rising from changes in weather patterns and/or climatic conditions produced by the Earth's state and behaviour. These typically span a longer period and/or occur at large scale (shifting precipitation patterns, rising average temperatures, long-term droughts, wildfires)
	GEO: Geohydrological	Addresses problems rising from mass transfer of water or solid earth, caused by the interaction of atmospheric conditions with land and ocean systems (floods, landslides, glacial melting, coastal erosion)
	ENV: Environmental	Addresses problems rising from degradation of ecosystem services and natural systems, posing threat to water, land, air, biodiversity, as well as crucial earth processes (deforestation, soil degradation, air pollution, desertification)
	BIO: Biological	Addresses problems rising from hazards of organic origin, affecting populations (human or animal), plants, crops, and/or endangered flora and fauna (vector-borne diseases, invasive species, animal epidemics)
Sector		<i>A sector is only assigned to a policy element if the link is explicitly stated or can be identified through the surrounding context</i>
	FOO: Food production	Addresses any of the following sub-sectors: agriculture/crop production; animal and livestock health and productivity; fishery yields and aquaculture production
	WAT: Water resources	Addresses any of the following sub-sectors: availability, accessibility, quality, of water; management of water resources
	HEA: Health and wellbeing	Addresses any of the following sub-sectors: air quality; heat; mental health; malnutrition; infectious diseases
	BUI: Built environment	Addresses any of the following sub-sectors: infrastructure, cities, settlements, communities, tourism, or economic sectors
	NAT: Natural environment	Addresses any of the following sub-sectors: terrestrial, freshwater, or ocean ecosystems; biodiversity; phenology

Appendix Table 2: Overview of the hazard and sector typology for annotating adaptation policy.

C Examples of premise-hypothesis pairs

A set of examples of the premise-hypothesis pairs, used to fine-tune the policy NLI agent, is provided in Appendix [Table 5](#).

D Fine-tuning details

The annotation model (SLM 1) was fully fine-tuned for 3 epochs using the *AdamW* optimizer with an effective batch size of 32 (per-device batch size - 4; gradient accumulation steps - 8) and a learning rate of $5e-5$. We used mixed-precision (FP16) training and gradient checkpointing to reduce required memory. The policy NLI agent and adaptation intentionality predictor were fine-tuned with similar

settings, except for the (per-device) batch size (16). Finally, the *ClimateBERT* model fine-tuned as the benchmark model for element type classification was fine-tuned with a lower learning rate ($5e-6$).

Fine-tuning was performed on a single *NVIDIA Tesla V100-SXM2-16GB* GPU, which has a rated maximum power consumption of 300 W. When operating at maximum power consumption continuously, the estimated power consumption per hour of fine-tuning equals 0.30 kWh. Considering that the fine-tuning processes ranged between 20 minutes and 2 hours for the SLMs in our framework, we roughly estimate the total power consumption of fine-tuning all three models to approximate 1 kWh.

Premise (policy element)	Class	Hypothesis (class description)	Label
promoting climate-resilient agricultural practices	Food production (sector)	the element addresses agriculture/crop production, animal and livestock health and productivity, fishery	1 (entailment)
At least 3 complete sets of specialised hazards maps & detail analysis in place	Sector-oriented (Goal subtype)	the element aims at achieving a desired state and/or solving a problem in a specific area of interest, such as a (social) group or region	0 (neutral)
provision of financial assistance to forest communities affected by climate change	Instrument-oriented (Goal subtype)	the element involves the introduction, change, or continuation of a policy instrument as a policy goal	1 (entailment)
Plan for Implementing Climate Change Science	Biological (hazard)	the element addresses problems rising from hazards of organic origin, affecting populations (human or animal), plants, crops, and/or endangered flora and fauna	0 (neutral)

Appendix Table 5: Examples of premise-hypothesis pairs used for fine-tuning a NLI-inspired classifier.

E Methods per annotation task

Here, we provide an overview of the detailed methodology for each of the tasks in the annotation process.

Task 1 (element extraction) – evaluating the accuracy of extracting the text sequences that represent a policy element comes with several challenges and can be approached from several angles and scales. To get a comprehensive understanding of the annotators’ sequence extraction accuracy, we introduce three metrics: token-level pooled F1 score (tok_{pool}), weighted sequence-level pooled F1 score (seq_{pool}), and weighted sequence-level macro-averaged F1 score ($\text{seq}_{\text{macro}}$).

Since the task involves free-text annotation, using only exact-boundary matches may lead to correct, partial matches being missed (Segura-Bedmar et al., 2013). Furthermore, different interpretations can lead to annotators splitting sequences differently, but a crucial part of establishing the evidence base is that the right parts of the text are extracted. The tok_{pool} metric, therefore, zooms out to the full set of policy elements per annotator, from which it creates a list comprising all tokens that are not a stop word (e.g., “the” or “at”). Following, the lists of tokens of annotator A and B are compared by counting tokens as true positive (token from list

B that also occurs in list A), false positive (token from list B that does not occur in list A), or false negative (token from list A that does not occur in list B) and computing the tok_{pool} F1 score.

While tok_{pool} helps to understand how well the annotators extract the right information, the sequence scores (seq_{pool} and $\text{seq}_{\text{macro}}$) measure the accuracy of the extracted policy elements as units. Considering that a boundary mismatch does not always mean that two sequences do not refer to the same entity (e.g., “the new national climate strategy of 2050” versus “national climate strategy of 2050”), we assess semantic similarity of two sequences (of different annotator on the same input chunk) if they have overlapping tokens. If the semantic similarity score exceeds 0.XX, they are considered to be referring to the same policy element, which allows for counting each policy element as a true positive, false positive, or false negative.

For the pooled variant (seq_{pool}), the F1 score is based on the total set of policy elements per annotator, indicating to what extent the extracted policy elements overlap. To this score, dense chunks (i.e., text blocks in which many policy elements are found) contribute more strongly compared to chunks in which only a few (or no) policy elements are identified. In contrast, the macro-averaged

score ($\text{seq}_{\text{macro}}$) takes the average of the F1 score per chunk, enabling an understanding of the extent to which the chunks are processed equally, and, e.g., correctly extracting zero elements from a chunk in which no relevant adaptation policy is found is rewarded. Here, the annotation accuracy of each chunk contributes to the score equally, regardless of the number of policy elements extracted.

The seq_{pool} and $\text{seq}_{\text{macro}}$ scores both use weights, which are determined by the adaptation intentionality scores of the policy elements. An element that intentionally addresses adaptation (e.g., building flood defence mechanisms) is considered to be more crucial not to miss out on compared to more indirect adaptation measures (e.g., climate awareness programmes). The total counts used as true positives, false positives, and false negatives are, therefore, the sums of the elements' intentionality scores, resulting in weighted F1 scores. The sum of true positives and false negatives in case of no overlapping policy elements in a single chunk, when computing seq_{pool} , forms an exception. In such cases, the F1 score will always equal zero, and there is no difference in score between, e.g., a single false negative with a low intentionality score as opposed to, e.g., five intentional policy elements being missed. We therefore introduce a custom formula to substitute the weighted sum of true positives to compute seq_{pool} :

$$TP_{\text{substitute}} = 1/(1+FN_{\text{weighted}}), \quad \text{if } TP = 0$$

$$FN_{\text{substitute}} = 1/(1+FP_{\text{weighted}}), \quad \text{if } FN = 0$$

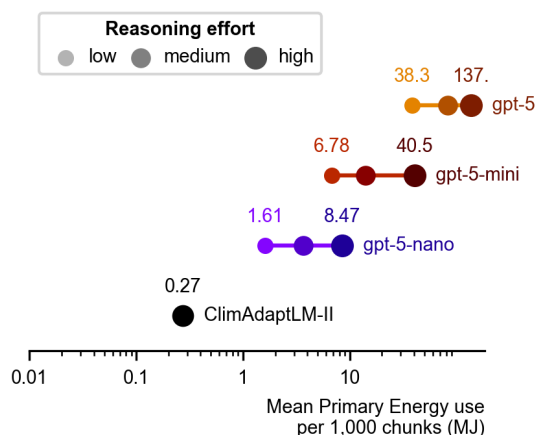
To illustrate, when annotator A did not identify any policy elements, and annotator B identified one element with an intentionality score of 0.4, the true positive substitute would equal 0.71, leading to a recall of 0.64. In case annotator B identified six elements with an average intentionality score of 0.75, however, the substitute equals 0.21, hence a recall of 0.04 – penalizing the missing annotation overlap more strongly. If both annotators identified no elements in a given chunk, we take an F1 score of 1 for that chunk.

Task 2-5 ((sub)type, hazard, and sector classification) – based on the matched policy elements arising from evaluation of the first task, the F1 scores of predicting the type (i.e., *Goal* or *Instrument*), subtype (e.g., process-oriented goal), hazards, and sectors are computed. F1 scores are reported per class.

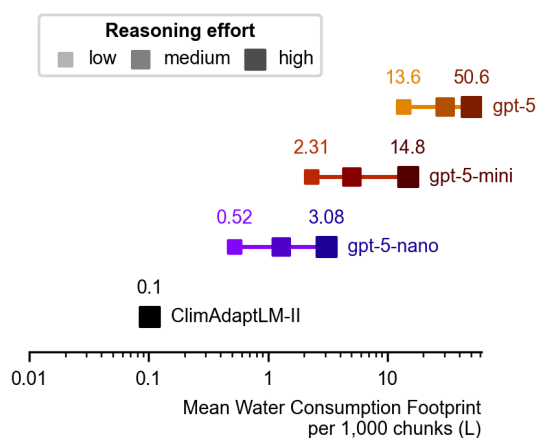
Task 6 (intentionality prediction) – to determine the direct accuracy of the regression model, we compute the Mean Squared Error (MSE). Additionally, the intentionality class labels assigned by the benchmark LLM (*gpt-5-mini*) are aligned to the intentionality scores of the SLM and human annotator.

F Environmental impact

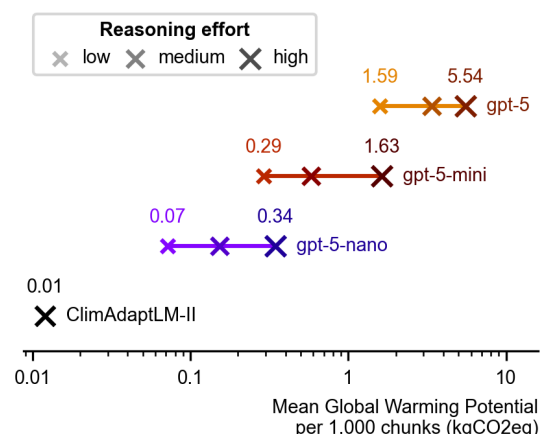
See Appendix Figure 1, Appendix Figure 2, Appendix Figure 3, and Appendix Figure 4.



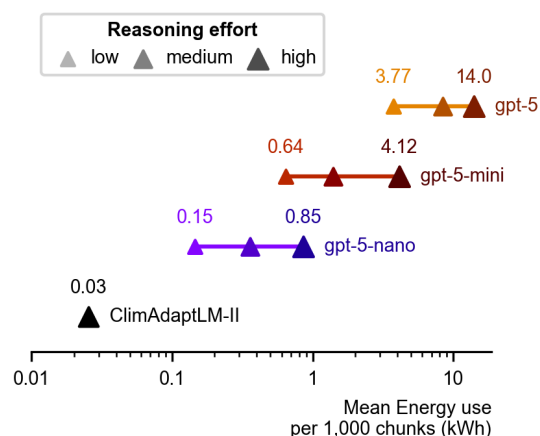
Appendix Figure 1: Estimated mean primary energy use per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.



Appendix Figure 2: Estimated mean water consumption per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.



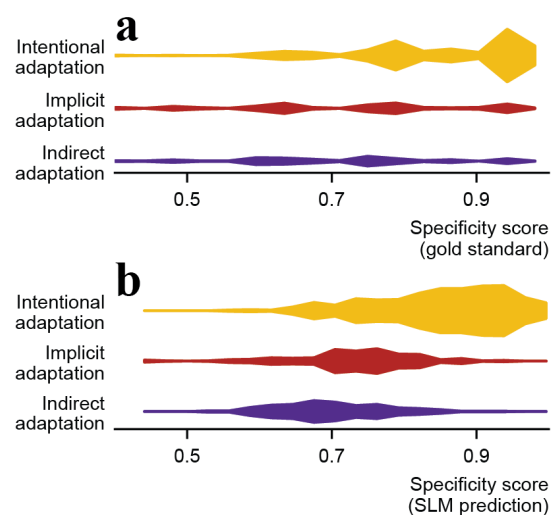
Appendix Figure 3: Estimated mean global warming potential per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.



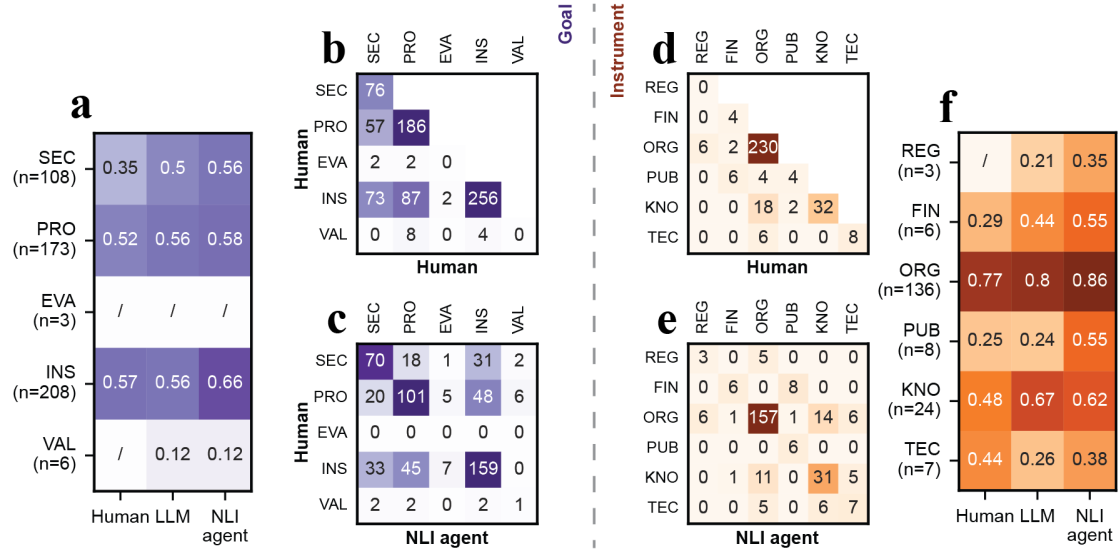
Appendix Figure 4: Estimated mean energy use per model when extracting policy elements from 1,000 chunks. The (logarithmic scale) x-axis shows the difference in primary energy use when changing the reasoning effort settings to low, medium, or high.

G Class agreement

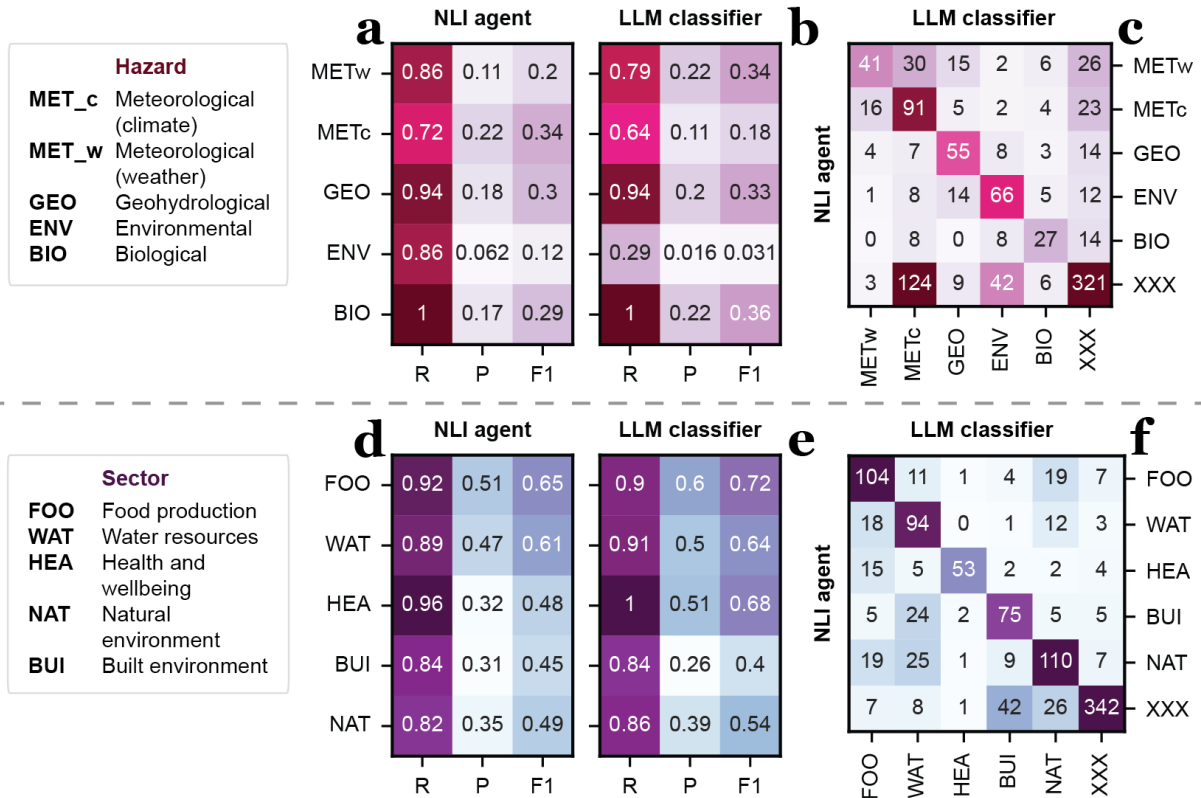
See Appendix Figure 5, Appendix Figure 6, and Appendix Figure 7.



Appendix Figure 7: Counts of intentionality class predictions by an LLM classifier per intentionality score interval. The filled areas show the total number of policy elements classified as the intentionality class (y axis) of each specificity score interval, determined by splitting the scores into 20 equal bins.



Appendix Figure 5: Accuracy and inter-annotator agreement of predicting subtypes (task 3). Subfigures (a) and (f) show the F1 scores per annotator (x axis), per class (y axis). The number of elements of that class in the test set is printed below the class labels (y axis). The confusion matrices show the confusion of two human annotators (b and d) and the NLI agent versus a human annotator (c and e) on the Goal (left) and Instrument (right) subtypes.



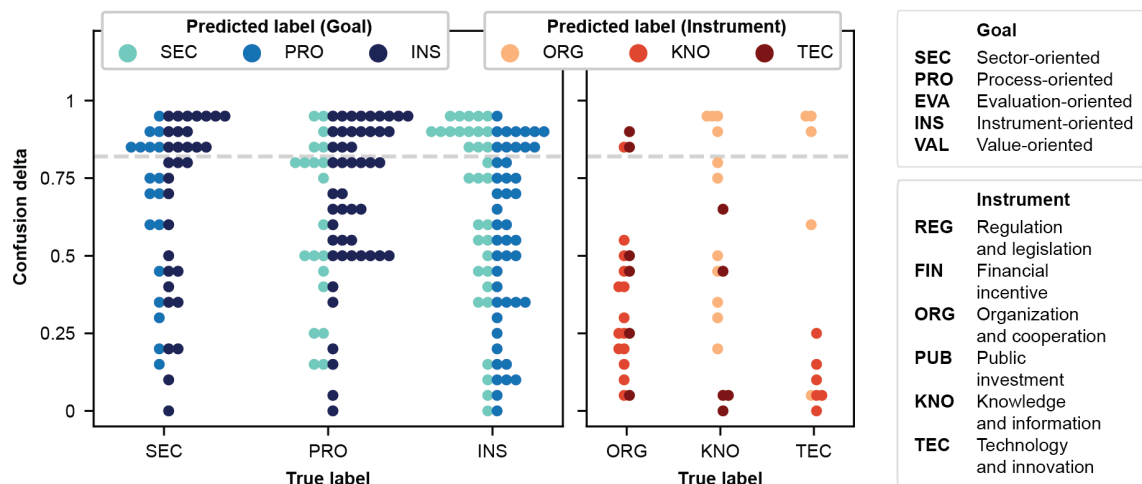
Appendix Figure 6: Accuracy of the policy NLI classifier and benchmark LLM on hazard and sector classes. Subfigures (a), (b), (d), and (e) show the recall, precision, and F1 score of the NLI agent (a and d) and benchmark LLM against the human-labelled gold standard on the hazard (a and b) and sector (d and e) classes. The confusion matrices in (c) and (f) show the agreement between the NLI agent and the benchmark LLM on the hazard (c) and sector (f) classes.

H Subtype mismatch analysis

To understand disagreement between annotators on class labels, we assessed substantial mismatches between the gold standard and predicted subtype (task 3) in detail. The selected mismatches are those where the confusion delta (i.e., the difference in probability of the gold standard and predicted label based on the NLI classifier) exceeds 0.8 (above the dashed lines in 8). For this process, we have prepared a dataset in which the two labels (gold standard are predicted) are randomly shuffled for review to account for possible bias. The review dataset consists of 79 annotated goals and 15 annotated instruments. On each policy element, we have indicated whether label A is correct, label B is

correct, or both apply, revealing the extent to which deviations from the gold standard are incorrect, human errors, or a result of ambiguity.

Among the 79 reviewed *Goal* elements, both classes were considered applicable in 10 cases, as opposed to 6 cases where none of the two labels were relevant according to the reviewer. Out of the remaining reviewed instances, the label given by the specialist SLM was chosen 39 times (62%) while the human annotator’s assigned label was considered correct 24 times (38%). Within the set of Instrument mismatches, the difference is even larger: only 2 of the human-assigned classes were rated as the correct label as opposed to 13 of the SLM-assigned classes (87%).



Appendix Figure 8: Subtype mismatches between gold standard (x axis) and predicted label. Each bubble represents a label mismatch, where the y location indicates the difference in model confidence on the predicted and true label.

I Model prompts

I.1 Element extraction (task 1)

Instruction

You are an expert in extracting climate policy elements from text. Your task is to retrieve text sequences that represent a policy element that relates to climate change from the input text. Make sure to return JSON objects that contain verbatim substrings and do not correct, modify or paraphrase the text in any way. Do not include punctuation at the end of the substring. Assign the most suitable policy type to each object.

Policy element

A policy element is a sequence of words that refers to a goal, instrument, or output that can be seen as a means to contribute to an overarching objective, in this case any element related to the **climate change** problem.

It might be that the provided text does not contain any policy elements. If so, return an empty array.

Inclusion criteria

1. Direct reference to climate impacts: there is an explicit mention of risks, hazards, or vulnerability linked to climate change (e.g., early warning systems for floods).
2. Focus on building resilience or vulnerability reduction: the element focuses on improving the ability of communities or (eco)systems to limit the impact (proactive) or recover (reactive) from expected climate impacts (e.g., objectives to reduce drought vulnerability in dry areas)
3. Indirect link: the element does not include a clear link to intentional adaptation or mitigation – but is known to indirectly contribute to, e.g., adaptive capacity, increased resilience, and/or reduced vulnerability (e.g., sustainable land use, climate research programs, climate subsidies for farmers)
4. Alignment with broader objectives: the element is mentioned to contribute to or support an overarching (e.g., international) objective through a specific goal or means (note: not the mention of the overarching objective itself)
5. Introduction of a new climate policy or inclusion of climate change in an existing policy. This also includes, e.g., programmes and projects targeted at climate change

Exclusion criteria

1. Unrelated or general development policy: the element describes actions that are not specifically linked to climate risks (e.g., building schools or reducing poverty). If it does focus on general climate change (e.g., prioritizing climate change in school programs) or is listed as a subitem of climate change policy (e.g., reducing poverty to increase resilience), however, it should be included (see item 3 of the inclusion criteria)
2. Disaster risk reduction elements that focus on short-term recovery or non-climate-related risks
3. Support to other countries: the retrieved policy elements should have a ****national**** focus
4. Overview of the contents of a document or policy: the element describes what another policy (document) consists of or targets, without providing a specific goal/instrument/output
5. Mention of the current document (title): if the text simply repeats the title or type of the current document, it should not be marked as a policy element

Policy domain

Each climate policy element is assigned one of the following climate policy domains:

1. "Adaptation": refers to the process of adjusting natural or human systems in response to actual or expected climatic changes and their effects, in order to moderate potential harm, exploit beneficial opportunities, or cope with unavoidable consequences
2. "Mitigation" : refers to efforts to reduce or prevent the emission of greenhouse gases (GHGs) into the atmosphere, or to enhance the removal of these gases through natural or technological sinks. The goal of

mitigation is to limit the magnitude or rate of long-term climate change itself, addressing its causes rather than its consequences

3. "General": these elements may target both adaptation and mitigation or be overarching elements (e.g., climate awareness)

JSON schema

```
{ "type": "array", "items": {
  "type": "object", "properties": {
    "policy_element": { "type": "string"
  }, "policy_domain": { "type":
    "string", "enum": ["Adaptation",
    "Mitigation", "General"] } }, "required":
    ["policy_element", "policy_domain"] } }
```

1.2 Element type classification (task 2)

Instruction

Your task is to classify adaptation policy elements as "Goal" or "Instrument". Only return "Goal" or "Instrument"; do not add any other text. Considering the following class definitions:

Goal

***Description**:** Quantitative or qualitative objective that a party aims to achieve, with the overarching aim of adapting to climate change. Involves a desired/aspired outcome or intended state/action

***Semantic indicators**:** Use of ambitious and/or future-oriented language ('to achieve', 'aiming to', 'target') or terms typically used for goals (e.g., 'increase', 'ensure', 'promote', 'support')

Instrument

***Description**:** The means to achieve the defined targets, typically in the form of a tool, mechanism, or strategy. Specifies how a goal can be achieved and can be written in present or past tense

***Semantic indicators**:** Mention of a specific tool, action, or intervention (e.g., 'regulation', 'program') or tangible action (e.g., 'build dykes')

I.3 Subtype classification (task 3)

Goal subtypes

Your task is to classify a climate change adaptation policy goal.

Pick the category (out of the 5 options) that best fits the provided goal. Return the three-letter code of the category without generating any other text.

Categories:

SEC* - Sector-oriented goal.

This category applies when the policy element aims at achieving a desired state and/or solving a problem in a specific area of interest, such as a (social) group or region. Related themes can be: accessibility and security of goods and services; well-being and prevention; development and mobility.

PRO* - Process-oriented goal.

This category applies when the policy element involves the organisational aspect of the policy, a process indirectly intended to address an issue. Related themes can be: cooperation; capacity-building; strategic planning.

EVA* - Evaluation-oriented goal.

This category applies when the policy element involves meeting a set evaluation criterion in a given sector. Related themes can be: quality; sustainability; effectiveness.

INS* - Instrument-oriented goal.

This category applies when the policy element involves the introduction, change, or continuation of a policy instrument as a policy goal. Related themes can be: regulation; investment; (re)organisation; knowledge and awareness.

VAL* - Value-oriented goal.

This category applies when the policy element aims at ensuring or advancing core constitutional values of democracies or political fundamentals. Related themes can be: justice and law; environmental protection; rights and equality.

Instrument subtypes

Your task is to classify a climate change adaptation policy instrument.

Pick the category (out of the 6 options) that best fits the provided instrument. Return the three-letter code of the category without generating any other text.

Categories:

REG* - Regulation and legislation instrument.

This category applies when the policy element uses the state's legal authority to order, forbid, or permit certain behaviour.

FIN* - Financial incentive instrument.

This category applies when the policy element uses taxes, subsidies, or tradable allowances to stimulate behaviour or activities.

ORG* - Organisation and cooperation instrument.

This category applies when the policy element involves the use of governmental resources through programmes, services, plans, and/or (voluntary) agreements.

PUB* - Public investment instrument.

This category applies when the policy element involves a strategic investment and/or the management thereof, done by public bodies.

KNO* - Knowledge and information instrument.

This category applies when the policy element involves research, education,

training, and/or knowledge sharing, focusing on knowledge generation and/or increasing awareness.

TEC* - Technology and innovation instrument.

This category applies when the policy element involves creation, advancement, and/or commercialization of new technologies or innovative systems or incentives.

I.4 Hazard classification (task 4)

Your task is to identify the hazard(s) addressed by a climate change adaptation policy element.

Pick 0, 1, or multiple categories (out of the 5 options) that fit the provided element.

Only select a category when the policy element ****explicitly**** addresses the hazard.

Return an array of the three-letter codes of the categories, without generating any other text.

If no hazard category fits the adaptation policy element, return an empty array ([]).

Categories:

MET_w* - Meteorological (weather) hazard.

This category applies when the policy element addresses problems rising from extreme weather events produced by the Earth's state and behaviour, occurring during a short period (minutes to days) and/or on small scale. Related themes can be: storms, extreme temperatures, heavy rain/snowfall, heatwaves, short-term droughts.

MET_c* - Meteorological (climate) hazard.

This category applies when the policy element addresses problems rising from changes in weather patterns and/or climatic conditions produced by the Earth's state and behaviour, spanning a longer period and/or occur at large scale. Related themes can be: shifting

precipitation patterns, rising average temperatures, long-term droughts, wildfires.

GEO* - Geohydrological hazard.

This category applies when the policy element addresses problems rising from mass transfer of water or solid earth, caused by the interaction of atmospheric conditions with land and ocean systems. Related themes can be: floods, landslides, glacial melting, coastal erosion.

ENV* - Environmental hazard.

This category applies when the policy element addresses problems rising from degradation of ecosystem services and natural systems, posing threat to water, land, air, biodiversity, as well as crucial earth processes. Related themes can be: deforestation, soil degradation, air pollution, desertification.

BIO* - Biological hazard.

This category applies when the policy element addresses problems rising from hazards of organic origin, affecting populations (human or animal), plants, crops, and/or endangered flora and fauna. Related themes can be: vector-borne diseases, invasive species, animal epidemics.

I.5 Sector classification (task 5)

Your task is to identify the sector(s) addressed by a climate change adaptation policy element.

Pick 0, 1, or multiple categories (out of the 5 options) that fit the provided element.

Only select a category when the policy element ****explicitly**** addresses the sector.

Return an array of the three-letter codes of the categories, without generating any other text.

If no sector category fits the adaptation policy element, return an empty array ([]).

Categories:

FOO - Food production sector.

This category applies when the policy element addresses agriculture/crop production, animal and livestock health and productivity, fishery yields, or aquaculture production.

WAT - Water resources sector.

This category applies when the policy element addresses availability, accessibility, quality, of water or management of water resources.

HEA - Health and wellbeing sector.

This category applies when the policy element addresses air quality, heat, mental health, malnutrition, or infectious diseases.

BUI - Built environment sector.

This category applies when the policy element addresses infrastructure, cities, settlements, communities, tourism, or economic sectors.

NAT - Natural environment sector.

This category applies when the

policy element addresses terrestrial, freshwater, or ocean ecosystems, biodiversity, or phenology.

I.6 Intentionality classification (task 6)

Your task is to classify adaptation policy elements into a ***intentionality*** class.

Based on the provided policy element, determine whether it intentionally (INT), implicitly (IMP), or indirectly (IND) addresses climate change adaptation.

Consider the following examples:

Intentional: flood defense mechanisms, resilience programmes

Implicit: climate research, disaster risk reduction

Indirect: sustainable development, climate awareness

Return only the three-letter code (INT, IMP, or IND) without returning any other text.