Transforming adaptation tracking: Benchmarking Transformer-based NLP approaches to retrieve adaptation-relevant information from climate policy text

Jetske Bonenkamp, Robbert Biesbroek, Ioannis Athanasiadis
Wageningen University and Research, Wageningen, The Netherlands
(jetske.bonenkamp, robbert.biesbroek, ioannis.athanasiadis)@wur.nl

Abstract

The voluminous, highly unstructured, and intersectoral nature of climate policy data resulted in increased calls for automated methods to retrieve information relevant to climate change adaptation. Collecting such information is crucial to establish a large-scale evidence base to monitor and evaluate current adaptation practices. Using a novel, hand-labelled dataset, we explored the potential of state-of-the-art Natural Language Processing methods and compared the performance of various Transformer-based solutions to classify text based on adaptation-relevance in both zero-shot and fine-tuned settings. We find that fine-tuned, encoder-only models, particularly those pre-trained on data from a related domain, are best suited to the task, outscoring zero-shot and rule-based approaches. Furthermore, our results show that text granularity played a crucial role in performance, with shorter text splits leading to decreased performance. Finally, we find that excluding records with below-moderate annotator confidence enhances model performance. These findings reveal key methodological considerations for automating and upscaling text classification in the climate change (adaptation) policy domain.

1 Introduction

The urgent need for climate change adaptation (referred to as 'adaptation' hereafter) has driven governments to formulate and implement ambitious policies and actions (Orlove, 2022). A comprehensive understanding of global adaptation progress, however, has remained absent. Despite conceptual proposals, no consistent, large-scale framework for tracking progress has been implemented to date (Magnan & Chalastani, 2019).

A key factor in this challenge is the abundance and unstructured nature of the relevant evidence, with adaptation information often being embedded in long climate policy documents. This hinders accessibility of relevant information to inform monitoring and evaluation, making identification of adaptation-relevant text essential for a tracking framework. The sheer volume of the text available, however, makes manual analysis infeasible, thus requiring an automated text classification approach. The field of Natural Language Processing (NLP) has shown great promise to contribute to adaptation tracking (Ford et al., 2016; Sietsma et al., 2024), but the multitude of approaches, setups, and data strategies that can potentially influence performance makes selecting the most suitable method challenging.

Rule-based approaches (e.g., keyword search) are most transparent and may achieve satisfactory results for non-complex topics, but their statistics-based successors are typically more accurate and stable (Li et al., 2022). For classification of short texts, early Deep Learning-based approaches continued this rising trend in accuracy, albeit with small margins – particularly when the dataset gets more imbalanced – and at the cost of computational efficiency (Shyrokykh et al., 2023).

More recently, the NLP field has shifted to the use of pre-trained models based on the Transformer architecture (Fields et al., 2024; Vaswani et al., 2017), of which encoder-only language models (ELMs) like BERT (Devlin et al., 2019) and large language models (LLMs) like GPT (Radford et al., 2019) are examples. For text classification tasks, state-of-the-art (SOTA) models have shown potential through three main approaches: (1) supervised fine-tuning of an ELM on a labelled dataset; (2) using an existing ELM fine-tuned on Natural Language Inference (NLI) for zero-shot classification, and; (3) prompting an advanced, general-purpose LLM to classify in a zero- or few-shot setting.

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 266–277
July 31, 2025 ©2025 Association for Computational Linguistics

The choice of approach and model, and their performance relative to more traditional NLP approaches, depends on numerous aspects. Prior research has shown that fine-tuned ELMs tend to outperform general-purpose LLMs on classification tasks when sufficient training data is available (Bucher & Martini, 2024), when the model is pre-trained on domain-relevant data (Dimitar et al., 2023), or when the task is of limited complexity (Yu et al., 2023). However, when training data is scarce or the text complexity requires advanced language understanding, LLMs may outperform fine-tuned ELMs (Yu et al., 2023), as well as traditional and NLI-based models (Z. Wang et al., 2023).

Other influential factors include text splitting strategy and inclusion threshold. Longer texts preserve context but pose challenges for SOTA models, as: (1) these are typically pre-trained on shorter texts (Fiok et al., 2021); (2) the models have difficulties with identifying information when text becomes more sparse (D'Cruz et al., 2024), and; (3) computation of Transformers scales quadratically with input length (Beltagy et al., 2020), making it challenging to determine the right text granularity and splitting strategy. Inclusion threshold refers to the extent to which a given text block must align with a label to belong to that class, potentially affecting model adaptability, and, therefore, performance. For zero-shot classification, setup choices like labels, task type (e.g., binary or multi-class), and prompt design (for LLMs) may also impact results.

To address the uncertainties discussed above, SOTA and traditional, automated text classification approaches are benchmarked against manually labelled climate policy texts. In addition, the impact of text granularity and inclusion thresholds is assessed. The aim is to identify the best method – i.e., the combination of approach, setup, and dataset variant (see Appendix A. for the nomenclature) getting closest to human-labelled, 'ground truth' examples – for extracting adaptation-relevant information from climate policy texts. By doing so, this work supports the creation of a global evidence base of adaptation progress.

2 Data

The main dataset comprises text extracted from national policy documents in the Climate Policy Radar (CPR) database[1] (Climate Policy Radar, n.d.), filtered to include only documents pre-labelled as adaptation-relevant and UNFCCC submissions, excluding mitigation-focused NIRs. A sample of 14 countries[2] (243 documents) was carefully selected to represent variety in climatic zones (WorldAtlas, 2023), developmental levels, number of available documents, and administrative language. All text was parsed from publicly available PDFs, transformed into Markdown format based on PDF layout, and non-English texts were translated via the Google Translate API using the default API settings (Han, n.d.). Subsets were created to evaluate the effects of text splitting and data cleaning strategies, as detailed in the following sections.

2.1 Chunking strategy

For assessing the effect of text granularity and context, three subsets of the main dataset were created. Each subset, referred to as 'dataset', uses a different strategy for splitting the texts into smaller blocks (i.e., chunking), as introduced below.

Dataset 1: Full chunks

First, the documents were split into text chunks of, on average, 3,186 characters and 10 paragraphs, using a Markdown-aware semantic splitter (*Semantic Text Splitter (API Documentation)*, n.d.). The chunks were sampled by document type, resulting in a set of 3,159 chunks, which were manually labelled by trained, graduate-level students and the authors of this paper. 24% of the dataset was labelled as relevant to adaptation. The inter-annotator agreement is 83%, which is considered acceptable. For the cases of disagreement between two annotators, the label with the highest confidence score was taken as the ground truth label. These confidence scores are further explained in section 2.2.

Dataset 2: Sub-chunks

To facilitate experimenting with variation in text splitting strategies, the text of dataset 1 was further split into sub-chunks of 500 to 800 characters so

[1] The data was retrieved from the database in June 2024. Documents added after this data are, therefore, not included in the dataset.

[2] Australia, Azerbaijan, Canada, Cyprus, Ecuador, Finland, Haiti, Iceland, Mexico, Nigeria, Sierra Leone, United Kingdom, Vanuatu, and Vietnam

that the chunks average approximately one paragraph and stay within the common NLP model limit of 512 tokens. For the full chunks previously labelled as ‘not adaptation’, the corresponding sub-chunks, totalling 13,132 items, were automatically assigned the same label. The remaining 5,356 ‘adaptation’ sub-chunks were re-labelled.

To address missing context caused by unclear coreferences, the sub-chunk experiments are conducted in two settings: one using the original sub-chunks without preprocessing (dataset 2a), and another applying coreference resolution (Elango, 2005) to replace unclear noun phrases (e.g., ‘the country’) with their parent entity (e.g., ‘Vietnam’) from outside the sub-chunk (dataset 2b). 71% of the sub-chunks retrieved from the relevant full chunks were re-labelled as adaptation-relevant, representing 20% of the full sample of 18,488 sub-chunks.

Dataset 3: Summarized chunks

To balance the advantages of shorter text but retaining crucial context, a third experimental dataset was created, in which the full chunks were summarized to single paragraphs using `bart-large-cnn`, a Transformer-based summarization model (Lewis et al., 2020). The automatically generated summaries were not evaluated at scale and thus may contain errors or inaccurate information. They are, therefore, solely used for the classification stage and not as actual adaptation evidence: the predicted labels are connected to the original, full chunks.

2.2 Data cleaning strategy

Besides the datasets resulting from applying the different chunking strategies, more dataset variants were created to assess differences in performance when applying different data cleaning strategies. Below, these strategies, each resulting in additional dataset variants[*], are introduced.

Confidence score threshold

A distinctive step was added in the annotation process. While labelling the chunks, the annotators specified a confidence score, indicating how sure they were about the assigned label (i.e., ‘adaptation’ or ‘not adaptation’) on a 0-100 scale. This score is used during evaluation to allow for assessing to what extent exclusion of chunks below a certain confidence score threshold (CST) affects performance. Figure 1 shows the distribution of the scores across the text chunks, indicating that for the majority of the chunks (i.e., >80%), the annotators were very confident (i.e., 80-100% certain).

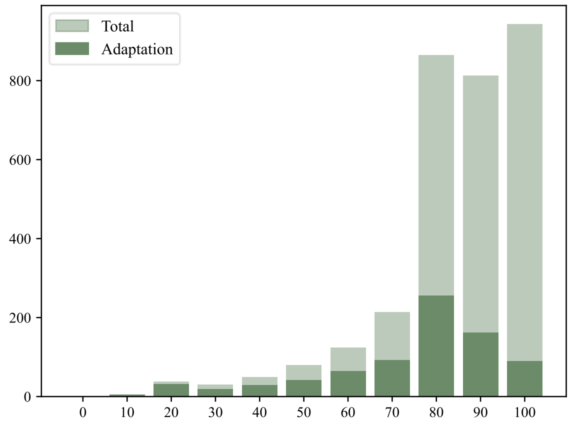

Figure 1: Distribution of confidence scores among 3,159 hand-labelled text chunks. The darker bars show the ratio of chunks labelled as ‘adaptation’.

Document type filter

When training classifiers, imbalanced data (i.e., uneven distribution of the classes) can cause difficulties for these models to correctly predict the right label, particularly for the under-represented class (Padurariu & Breaban, 2019). Since the document types of the full documents retrieved from CPR are known, an analysis of the class distribution per document type, based on the labels of dataset 2, revealed that there were multiple document types with a sub-chunk relevance ratio of 4% or lower[3]. Removing all chunks of these low-relevance document types would increase the initial (i.e., with no CST applied) ratio of adaptation-relevant chunks from 24% to 29% (dataset 1 and 3) and from 20% to 28% (dataset 2). The size of the datasets reduces from 3,159 to 2,572 (dataset 1 and 3) and from 18,488 to 13,948 (dataset 2) when applying this document type filter. To assess whether this increased balance, despite the decreased size of the training data, leads to improved performance that compensates for potentially missed relevant data, both strategies are applied and evaluated in combination with all datasets introduced in section 2.1. The resulting

[3] Decision and Plan, Regulation, Vision, Roadmap, Constitution, Act, Long-Term Low-Emission Development Strategy, and Biennial Update Report

[*] See Appendix A. for the nomenclature

dataset variants are referred to as *unfiltered* (i.e., all document types included) and *filtered* (i.e., low-relevance document types removed).

3 Methodology

Four main approaches are benchmarked against the dataset variants presented in section 2. Each approach and the corresponding sub-methods (i.e., models, queries, tasks, and/or prompts) are introduced in below sections. The confidence scores of the labels and the two filtering strategies (see section 2.2) are used to prepare ten variants of each dataset, each corresponding to a CST value of 0 (i.e., original labels maintained), 50, 60, 70, or 80 combined with *filtered* or *unfiltered* as the data cleaning strategy. For each CST iteratively, the items with label 'adaptation' but a score lower than the CST are excluded from the dataset. For each dataset variant, random splits are created, where 15% is used for evaluation, and, where applicable, 70% for training and 15% for validation. Performance is evaluated by computing precision, recall, and F1-score compared to the human-coded labels. Additional criteria, such as computational cost, are also noted during the final evaluation.

3.1 Rule-based classification (RBC)

The rule-based pattern matching technique is arguably the simplest approach evaluated, querying for (sets of) keywords to classify the chunks. Three different queries are applied. The first is a baseline query, focusing on the text sequence 'adapt' only. The second is theory-based, following the concept of adaptation (Orlove, 2022). The third query is data-driven, following the prominent topics in the data labelled as relevant, determined by applying a topic model. The queried topics were additionally filtered to exclude the topics that occurred in more than 200 of the chunks that were labelled as 'not adaptation' (e.g., 'Paris Agreement'). The resulting queries can be found in table 1.

3.2 Natural Language Inference (NLI)

Four NLI-based zero-shot classifiers are evaluated for identifying adaptation-relevant text: **`deberta-small-long-nli`** (Sileo, 2024), **`bart-large-mnli`** (Facebook, 2024), **`deberta-v3-large-zero-shot-v2.0`** (Laurer et al., 2024), and **`nli-MiniLM2-L6-H768`** (W. Wang et al., 2021). NLI models leverage their understanding of language obtained through pre-training to determine whether a hypothesis (label) is true given a premise (text) (Laurer et al., 2024). The model selection is based on compatibility with longer texts, model transparency, and reported performance in prior work. For each model, different tasks were evaluated, adding variety in used labels and task type (i.e., binary versus multi-class). An overview of the different tasks can be found in Appendix B. . For the 'multi-class' task type, where the model is asked to assign scores to multiple labels rather than a binary judgment about only the presence of an adaptation-related label, the additional label (i.e., 'mitigation') was ignored during evaluation, and the experiments were repeated with different thresholds for the adaptation label.

3.3 Fine-tuned encoder-only models (FEM)

Four models were selected to be fine-tuned for the classification task. They were chosen to include both general-purpose and domain-specific models, taking into account important criteria such as context length compatibility, pre-training data characteristics, and model parameters. The first one is the general-purpose model **`distilroberta-`**

Table 1: Overview of search queries

Title	Simplified expression
Baseline query	adapt[a-z]*
Theory-based query	adapt[a-z]* OR ((decreas[a-z]+ OR reduc[a-z]+ OR mitigat[a-z]+ OR avoid[a-z]*) NEAR (impact OR vulnerab[a-z]+ OR hazard OR exposure OR risk)) OR ((increas[a-z]+ OR improv[a-z]+ OR enhanc[a-z]+ OR build) NEAR resilien[a-z]+)
Data-driven query	((climat[a-z]+)? (change)? adapt[a-z]+) OR ((natural)? disaster NEAR (prevent[a-z]* OR control OR respons[a-z]+)) OR (risk NEAR (reduc[a-z]+ OR manag[a-z]+)) OR ((negative)? climat[a-z]+ NEAR impacts?) OR (climat[a-z]+ NEAR respons[a-z]+) OR ((sea level) NEAR rise) OR (capacit[a-z]+ NEAR build[a-z]*) OR ((climat[a-z]+ OR (fast start)) NEAR financ[a-z]+) OR ((early) warning NEAR system) OR (environment[a-z]* NEAR protect[a-z]*) OR (natural NEAR resource[a-z]+ NEAR manag[a-z]+)

Table 2: Overview of prompts used for LLM-based classification

ID	Prompt
P1 (concise)	Classify the following climate policy text chunk as "Adaptation" or "Not adaptation". Do not include any text other than the label.
P2 (specific)	Your task is to categorize text chunks as "Adaptation" or "Not adaptation". If the text contains any information about climate change adaptation policy, categorize it as "Adaptation". If not, for example when it only contains information about mitigation, categorize it as "Not adaptation". Do not include any text other than the label.

base (Sanh et al., 2019). The second is a climate domain-specific model from the ClimateBERT family, namely distilroberta-base-climate-f (Webersinke et al., 2022). The third model evaluated is legal-bert-small-uncased (Chalkidis et al., 2020), a model tuned to the legal domain, and the final one involves a model trained for understanding of environmental texts, namely EnvironmentalBERT-base (Schimanski et al., 2024).

All listed models are iteratively fine-tuned on the different dataset variants to assess the performance of the models themselves, as well as the impact of tuning the training data on the results.

3.4 Large Language Models (LLMs)

State-of-the-art LLMs are prompted to assign a binary label (i.e., 'adaptation' or 'not adaptation') to the text chunks they are provided with. Here, the experiments are conducted with OpenAI's GPT-4o (OpenAI et al., 2024) and the 8 billion parameter version of Llama 3.1 (Touvron et al., 2023). Table 2 provides an overview of the two prompts used. Given the length of the chunks combined with the abundance of the dataset, only zero-shot prompting techniques were included in the experiments: providing examples in the prompt (i.e., few-shot learning) would require excessive computational resources (Sahoo et al., 2025). Two prompt variations were applied, were the first one (P1) only provides the task to the model, and the second (P2) elaborates further on the label definitions.

The prompts were carefully composed to vary in conciseness (P1) and specificity (P2), following common prompt engineering principles (Geroimenko, 2025). This experiment is intended to bring insights into how extending the prompt with additional context information and elaborated instructions, thereby limiting conciseness, affects performance.

4 Results

The variations in approaches, setups, and dataset variants resulted in 791 different methods*. A complete overview of the evaluation scores of each method are accessible via the Git repository of this paper[4]. In the following sections, a selection of the most noteworthy results is presented.

4.1 Approaches

For each approach introduced in section 3, the two best methods based on F1 score and the best method based on recall are plotted in figure 2. The bars show the distribution of true positives (darker green), false positives (pastel green), and false negatives (orange). In below subsections, the results of each approach are discussed.

Rule-based classification (RBC)

The results of the RBC experiments reveal, as can be obtained from figure 2, that both the theory-based and data-driven queries outscored the baseline query on the recall metric. This indicates that these setups* excel at correctly identifying the largest ratio of relevant chunks. This increase, however, negatively affects precision, as the baseline query (i.e., only searching for the word 'adaptation') shows better results at limiting the number of irrelevant items being predicted as relevant. Overall, the data-driven query mainly outperforms the theory-driven one on recall.

Zero-shot classification (NLI/LLM)

The results of the two zero-shot approaches (see the bars of NLI and LLM in figure 2) show that the instructed LLMs provide better scores compared to the NLI-based models. Although the *BART-large*

[4] git.wur.nl/bonen003/transforming-adaptation-tracking

* See Appendix A. for the nomenclature

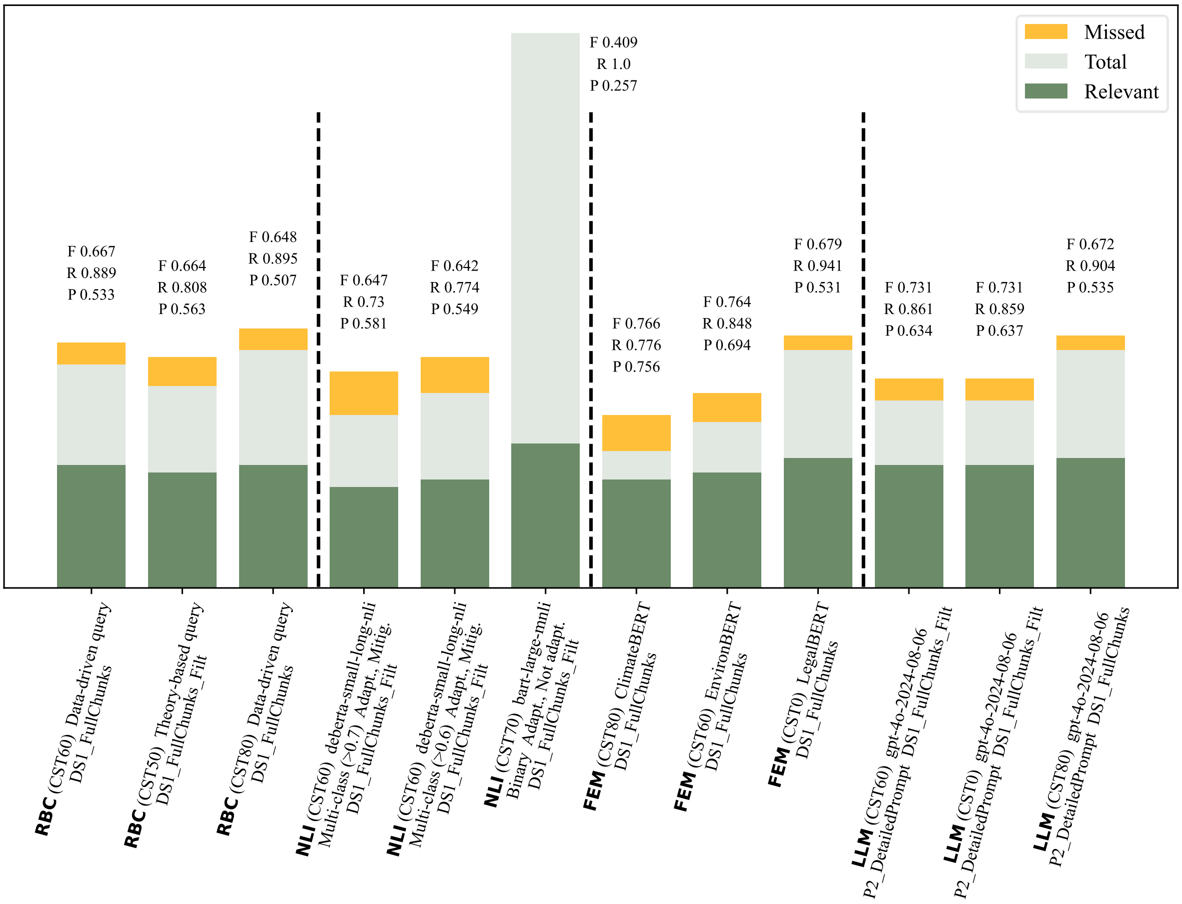

Figure 2: Selection of results for the four approaches, namely Rule-Based Classification (RBC), Natural Language Inference (NLI), Large Language Models (LLM) and Fine-tuned Encoder Models (FEM). For each approach, the two top-performing models based on F1 score (F) and the top one according to recall (R) were selected. Precision is also reported (P).

model achieves a perfect recall, it incorrectly classifies most of the irrelevant items as adaptation-relevant. Among the two LLMs evaluated, *GPT-4o* outperforms *Llama* on all occasions, being particularly well-capable of identifying relevant chunks. For *GPT-4o*, the specific prompt (P2) shows an increase in recall, although at the cost of precision.

Fine-tuned models (FEM)

The FEM experiment results show that three different domain-specific models occur among the three top scoring methods (see figure 2). The methods using *ClimateBERT* and *Environmental-BERT* achieve the best F1-score, indicating good capability of balancing inclusion of relevant items and exclusion of irrelevant items, with the differences mainly found in the balance between recall and precision. The *LegalBERT*-based method excels at recall, predicting 94.1% of the adaptation-relevant items as such.

4.2 Annotator confidence

For analysis of the effects of applying CSTs on the training and evaluation data, the CSTs were clustered into low (i.e., all labelled data included), medium (i.e., all items labelled with a score of 50% or lower excluded), and high (i.e., all items with a score of 70% or lower excluded). The bars in figure 3 show the mean evaluation scores of the best five methods per approach based on F1 score. Here, it becomes clear that applying a CST affects performance for all approaches, as a low CST yields the lowest scores in all four cases. For the traditional and zero-shot approaches, a higher CST positively affects the ratio of items correctly identified, whereas it mainly results in increased precision (i.e., the ratio of non-relevant items incorrectly predicted as relevant) for the fine-tuned models.

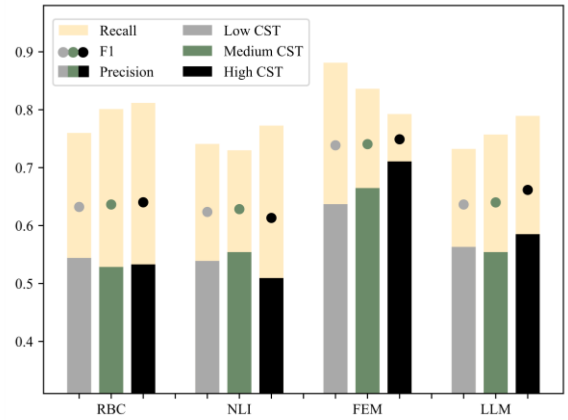

Figure 3: Results of applying a low, medium, or high confidence score threshold (**CST**) on the dataset. The average performance metrics of the top **five** methods per approach, ranked by F1-score, are plotted.

In the prompt variations used for classification with *GPT-4o*, it is observed that the prompt resulting in the best scores depends on what CST is applied. Although recall increases in all cases, the specific prompt (P2) lead to degraded performance when no CST was applied (i.e., CST0) or when a high threshold was used (i.e., CST80). For this zero-shot approach, a medium CST combined with a specific prompt (P2) results in the best scores, excelling on both recall and precision (see figure 4).

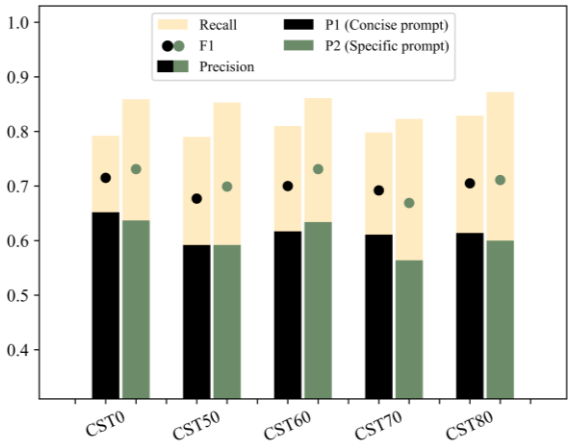

Figure 4: Results of classification with GPT-4o. The bars show F1-score, recall, and precision for each confidence score threshold (CST) and compares the scores of using a **concise** versus a **specific** prompt.

4.3 Chunking strategy

A comparison of the overall performance of each approach on the different datasets (see figure 5) shows that the best balance between maximized true positives and minimized false positives is achieved with the dataset of full-length text chunks (dataset 1). The models fine-tuned and/or evaluated on this dataset particularly excel on recall. Although the margins vary, the summarized dataset (dataset 3) resulted in the lowest F1 score for all approaches. The methods in non-fine-tuned settings do show an increase in precision for this dataset compared to the others.

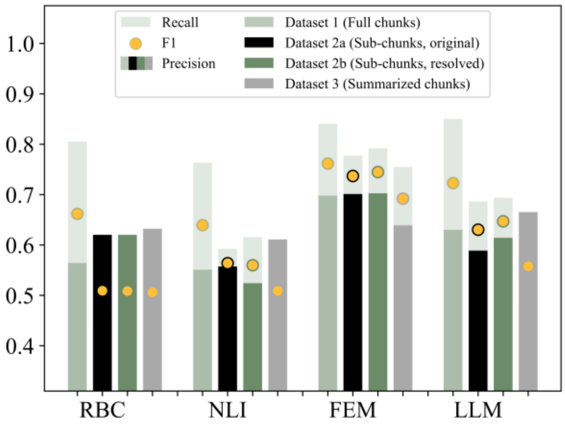

Figure 5: Results of evaluating classification with four different **datasets**: dataset 1 (full chunks), dataset 2a (sub- chunks), dataset 2b (sub-chunks with resolved coreferences), and dataset 3 (summarized chunks). The average performance metrics of the top **five** sub-methods per approach, ranked by F1-score, are plotted.

Comparing the two versions of the sub-chunk dataset, no major differences in performance between the original dataset (2a) and the one with resolved coreferences (2b) are observed. For the FEM approach, coreference resolution shows a slight increase in evaluation scores. However, in most other cases, as figure 5 also indicates, the models become less capable of identifying relevant items, hence a decrease in recall.

4.4 Data cleaning strategy

The results plotted in figure 6 reveal that adding a document type filter positively affects the F1 score for all approaches, except for the fine-tuned models. For this FEM approach, the results show that the data strategy (i.e. filtered versus unfiltered on document type) that outscores the other varies per model and CST. This is expected, as applying the filter lead to a more balanced dataset, typically improving classification performance, but the size of the training dataset decreases, meaning the model has less examples to learn from. Of all experiments conducted overall, both strategies occur in the top 20 (sorted by F1 score). The absolute numbers of a confusion matrix of the best-scoring variant of each filtering strategy[5] (see Appendix C.) also indicate that there is no clear outperforming data strategy*. The most suitable choice depends on various design choices, as further explored in the following section.

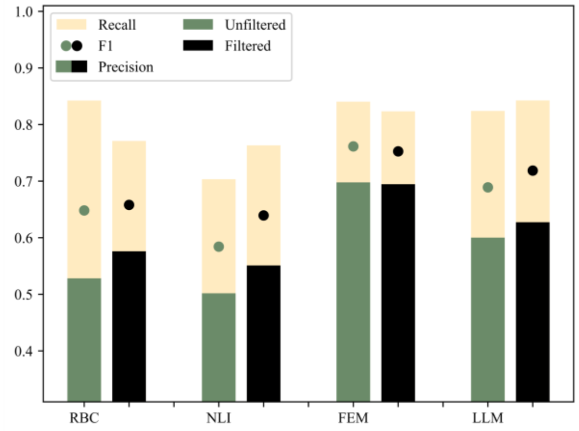

Figure 6: Results of evaluating classification with two different data strategies: with and without an applied **document type filter**. The average performance metrics of the top **five** methods per approach, ranked by F1-score, are plotted.

4.5 Overall comparison

In this research, recall is prioritized over precision, meaning that the 'best' method is not purely determined based on F1 score. Setting a precision threshold of 0.66 and sorting the results on recall leads to a set of four FEM methods considered most suitable to the task, each with its own strength. An overview of these methods, including their results, is provided in table 3. All models in this selection are fine-tuned and applied on/to the dataset of full chunks (dataset 1).

Table 3: Overview of four selected methods, referring to models fine-tuned on specific dataset variants

Model	**Var.**	**Prec.**	**Rec.**	**Comp. Cost**
ClimateBERT	CST70, Filtered	0.673	0.871	-
ClimateBERT	CST60, Unfilt.	0.672	0.869	-/o
LegalBERT	CST0, Filtered	0.671	0.855	-
EnvironBERT	CST60, Unfilt.	0.694	0.848	o

This selection shows that there are multiple methods* that lead to satisfactory results. The selected methods show comparable performance, from where it is obtained that the two models fine-tuned and evaluated on a filtered dataset variant are most computationally efficient (measured by duration of the fine-tuning process), the *EnvironmentalBERT*-based method excels on precision and F1 score, and the first *ClimateBERT* model achieves the best recall.

5 Discussion

The main objective of this paper was to determine the best classification method to identify adaptation-relevant text chunks in large and unstructured climate policy documents. The results reveal that each approach comes with its own strengths and weaknesses, but domain-specific models fine-tuned on a labelled dataset showed the best balance between ratio of correctly identified, relevant items and minimized presence of irrelevant items among those predicted as being relevant. With F1 scores of 0.759, 0.758, 0.752, and 0.764 respectively, four fine-tuned models (listed in table 3), including three different base models and multiple dataset variants, have proven their potential for identifying relevant information needed to track adaptation globally. These findings align with those of Bucher & Martini (2024), i.e., that fine-tuned models outperform LLMs when sufficient training data is available, and those of Dimitar et al. (2023), i.e., that better scores are achieved when such models have been pre-trained and/or previously fine-tuned on domain-specific data. In this research, where the labelled data has been created, this supervised FEM approach is considered most suitable, as it outperforms the benchmarked RBC and NLI approaches by large margins and shows small performance advancements over the top-scoring zero-shot method with *GPT-4o*. As the differences with the latter are relatively minor, however (i.e., an F1 difference of 0.07), LLM-based zero-shot classification has also demonstrated its potential. Especially in future cases, when the resources to (re-)create a labelled dataset are limited, this approach may be a valid alternative. The results have shown, however, that the chosen CST and

[5] Determined by setting a minimum precision of 0.66, then sorting by recall (descending)

* See Appendix A. for the nomenclature

prompt design can majorly affect the performance of LLMs as classifiers, making this approach less reliable and robust when there is no labelled dataset available for validation.

What specific FEM and dataset variant should be selected, however, depends on prioritization of trade-offs. First, what CST is applied should be taken into account for the final choice of method. A low CST means that even chunks that are somewhat relevant will be included in the eventual dataset, which limits the possibility of missing out on potentially relevant information. The relatedness of the text in the final dataset, however, likely improves when its content is relevant to adaptation with high confidence. Using a medium CST (i.e., 50-70) is, therefore, preferred, as this balances out these (dis)advantages.

In contrast to prior studies which suggest that performance of Transformer-based models typically improves when text is relatively short and consistent, the results show that for all approaches, the best scores were achieved using the full chunk dataset. Applying coreference resolution to the short text splits did not solve the 'missing context' problem, showing negligible differences, nor did automated summarization overcome the challenge of dealing with longer text lengths. This emphasizes the importance of context information, which likely connects to adaptation's conceptual indistinctness described by Dupuis & Biesbroek (2013). Using the full chunks (dataset 1) is, therefore, preferred here. Determining whether a document type filter should be applied, however, turned out a greater challenge. The main advantage of filtering the dataset on document type is that it limits the size (by more than 20%) of the dataset, positively affecting computational cost, and that it improves class balance. This, depending on the sub-method, results in an increased recall, compensating for the small ratio of relevant chunks that are missed out on by applying the filter.

Determining the overall best method requires an optimal trade-off between precision and recall. Although capturing all relevant information is crucial, ensuring sufficient precision to minimize the presence of irrelevant information and, with that, improve the quality of the evidence base, should not be ignored. Therefore, a precision minimum of 0.66 was set, after which the results were ranked by recall. In addition, computational cost also plays a role in determining the optimal method. For establishing the adaptation evidence base, therefore, the *ClimateBERT* model fine-tuned on the filtered dataset with a CST of 70 (see table 3) is considered the most appropriate choice.

These findings underscore the potential of state-of-the-art NLP methods to narrow down relevant policy information at large scale, which may also be interesting to explore in other (policy) domains. Other suggested future research directions involve successive steps in establishing an NLP-driven adaptation tracking framework by, e.g., further unpacking and structuring the unstructured climate policy texts by identifying and categorizing adaptation-specific (policy) elements.

6 Conclusion

This work has revealed important methodological considerations for classification of adaptation policy texts. For an automated framework for identifying relevant information, with the aim of creating a dataset of adaptation policy and, with that, increasing accessibility of information needed to track progress, a fine-tuned *ClimateBERT* model has shown optimal performance. This method ensures a sufficient balance between correctly identified text, minimized missed items, and maximization of irrelevant items filtered out. To boost performance, label confidence should be taken into account during manual labelling. Following, items labelled with a confidence score of less than 60% should be excluded. Also, documents should be filtered to include only those that are known to contain adaptation-relevant information and should be split based on Markdown structure and semantic meaning, with an average of 10 paragraphs per splits. The exact length is determined by the semantic splitter, ideally with a range of 2,000-8,000 characters.

Limitations

The discussed work comes with several limitations. First, the text chunks were automatically parsed from the original PDFs and non-English text was machine-translated. The data may, therefore, contain parsing and/or translation errors, potentially affecting the results. Second, relevance labels and confidence scores were assigned by human annotators, making them exposed to subjectivity. This was also observed in the inter-annotator agreement, where the annotators disagreed in 17% of the cases. Considering this, despite extensive training, the labels and scores may not always reflect true certainty, highlighting the ambiguity of the classification task and the challenge of aligning AI predictions with human judgment. Third, the automatically generated summaries (dataset 3) were not extensively reviewed and no alternative methods or models for summarization were explored, limiting comprehensive assessment of the potential of this approach. Fourth, only two prompt variations were evaluated, which were based on prompt engineering principles (Geroimenko, 2025) and may not reflect the full potential of the zero-shot LLM approach.

Following these limitations, future work should enhance validity by, e.g., delving further into annotation consistency, evaluating alternative summarization models, and full-scale evaluation of more than these two prompt variations to assess whether the practical results align with prompt design theory.

Acknowledgements

We thank Demi Ursem, Denatalie Chrisdameria, Suze Zander, and Robin Blersch for their valuable assistance in hand-labelling the datasets used in this research. In addition, we thank the anonymous reviewers for their useful feedback on an earlier version of this manuscript. Finally, JB and RB would like to acknowledge funding support from the Dutch Research Council (NWO VI.Vidi.211.132).

References

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer* (arXiv:2004.05150). arXiv. https://doi.org/10.48550/arXiv.2004.05150

Bucher, M. J. J., & Martini, M. (2024). *Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification* (arXiv:2406.08660). arXiv. https://doi.org/10.48550/arXiv.2406.08660

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). *LEGAL-BERT: The Muppets straight out of Law School* (arXiv:2010.02559). arXiv. https://doi.org/10.48550/arXiv.2010.02559

Climate Policy Radar. (n.d.). *Climate Policy Radar | AI for climate law and policy research*. Climate Policy Radar. Retrieved 30 May 2024, from https://www.climatepolicyradar.org/

D'Cruz, C., Bereder, J.-M., Precioso, F., & Riveill, M. (2024). *Domain-specific long text classification from sparse relevant information* (arXiv:2408.13253). arXiv. https://doi.org/10.48550/arXiv.2408.13253

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dimitar, T., Gorgi, L., Ljubomir, C., & Irena, V. (2023). Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks. *E3S Web of Conferences*, *436*, 02004. https://doi.org/10.1051/e3sconf/202343602004

Dupuis, J., & Biesbroek, R. (2013). Comparing apples and oranges: The dependent variable problem in comparing and evaluating climate change adaptation policies. *Global Environmental Change*, *23*(6), 1476–1487. https://doi.org/10.1016/j.gloenvcha.2013.07.022

Elango, P. (2005). *Coreference Resolution: A Survey*.

Facebook. (2024, January 4). *Model card for bart-large-mnli*. Hugging Face. https://huggingface.co/facebook/bart-large-mnli

Fields, J., Chovanec, K., & Madiraju, P. (2024). A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe? *IEEE Access*, *12*, 6518–6531. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3349952

Fiok, K., Karwowski, W., Gutierrez-Franco, E., Davahli, M. R., Wilamowski, M., Ahram, T., Al-Juaid, A., & Zurada, J. (2021). Text Guide: Improving the Quality of Long Text Classification by a Text Selection Method Based on Feature Importance. *IEEE Access*, *9*, 105439–105450. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3099758

Ford, J. D., Tilleard, S. E., Berrang-Ford, L., Araos, M., Biesbroek, R., Lesnikowski, A. C., MacDonald, G. K., Hsu, A., Chen, C., & Bizikova, L. (2016). Big data has big potential for applications to climate change adaptation. *Proceedings of the National Academy of Sciences*, *113*(39), 10729–10732. https://doi.org/10.1073/pnas.1614023113

Geroimenko, V. (2025). Key Principles of Good Prompt Design. In V. Geroimenko (Ed.), *The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks* (pp. 17–36). Springer Nature

Switzerland. https://doi.org/10.1007/978-3-031-86206-9_2

Han, S. (n.d.). *googletrans: Free Google Translate API for Python. Translates totally free of charge.* (Version 3.0.0) [Python; MacOS :: MacOS X, Microsoft :: Windows, POSIX].

Laurer, M., Atteveldt, W. van, Casas, A., & Welbers, K. (2024). *Building Efficient Universal Classifiers with Natural Language Inference* (arXiv:2312.17543). arXiv. https://doi.org/10.48550/arXiv.2312.17543

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol.*, *13*(2), 31:1-31:41. https://doi.org/10.1145/3495162

Magnan, A. K., & Chalastani, V. I. (2019). *Towards a Global Adaptation Progress Tracker: First thoughts*. Institute for Sustainable Development and International Relations (IDDRI). https://www.jstor.org/stable/resrep52383

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., … Malkov, Y. (2024). *GPT-4o System Card* (arXiv:2410.21276). arXiv. https://doi.org/10.48550/arXiv.2410.21276

Orlove, B. (2022). The Concept of Adaptation. *Annual Review of Environment and Resources*, *47*(1), 535–581. https://doi.org/10.1146/annurev-environ-112320-095719

Padurariu, C., & Breaban, M. E. (2019). Dealing with Data Imbalance in Text Classification. *Procedia Computer Science*, *159*, 736–745. https://doi.org/10.1016/j.procs.2019.09.229

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications* (arXiv:2402.07927). arXiv. https://doi.org/10.48550/arXiv.2402.07927

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019, October 2). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv.Org. https://arxiv.org/abs/1910.01108v4

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, *61*, 104979. https://doi.org/10.1016/j.frl.2024.104979

Semantic text splitter (API documentation). (n.d.). Retrieved 19 December 2024, from https://semantic-text-splitter.readthedocs.io/en/stable/semantic_text_splitter.html

Shyrokykh, K., Girnyk, M., & Dellmuth, L. (2023). Short text classification with machine learning in the social sciences: The case of climate change on Twitter. *PLOS ONE*, *18*(9), e0290762. https://doi.org/10.1371/journal.pone.0290762

Sietsma, A. J., Ford, J. D., & Minx, J. C. (2024). The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, *14*(1), 31–39. https://doi.org/10.1038/s41558-023-01890-3

Sileo, D. (2024). tasksource: A Large Collection of NLP tasks with a Structured Dataset Preprocessing Framework. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 15655–15684). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1361

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, W., Bao, H., Huang, S., Dong, L., & Wei, F. (2021). *MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers* (arXiv:2012.15828). arXiv. https://doi.org/10.48550/arXiv.2012.15828

Wang, Z., Pang, Y., & Lin, Y. (2023). *Large Language Models Are Zero-Shot Text Classifiers* (arXiv:2312.01044). arXiv. https://doi.org/10.48550/arXiv.2312.01044

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). *ClimateBert: A Pretrained Language Model for Climate-Related Text* (arXiv:2110.12010). arXiv. http://arxiv.org/abs/2110.12010

WorldAtlas. (2023, April 13). *The Climate Zones Of The World*. WorldAtlas. https://www.worldatlas.com/climate/the-climate-zones-of-the-world.html

Yu, H., Yang, Z., Pelrine, K., Godbout, J. F., & Rabbany, R. (2023). *Open, Closed, or Small Language Models for Text Classification?* (arXiv:2308.10092). arXiv. https://doi.org/10.48550/arXiv.2308.10092

Appendices

Appendix A. Table of nomenclature

For a clear overview of the terms used throughout the paper, one can refer to table 4. Any combination of the different levels (e.g., LLM-based classification with GPT-4o on the filtered variant of dataset 1) is referred to as a *method*. Any combination of levels 2 up to and including 5, for a given approach, is called a *sub-method*.

Appendix B. NLI task types

Table 5: Overview of tasks included in the NLI-based classification experiments

Labels	Task type
Adaptation	Binary
Adaptation policy	Binary
Climate change adaptation	Binary
Adaptation, Mitigation	Multi-class (>0.5) Multi-class (>0.6) Multi-class (>0.7)

Appendix C. Confusion matrix data strategy

Table 6: Confusion matrix of classification results on the test set in absolute numbers. Each cell shows the results of the filtered (L) versus unfiltered (R) data strategy.

		Predicted label	
		A	**NA**
True label	**A**	83 / **87**	**11** / 13
	NA	**40** / 42	188 / 301

Table 4: Table of nomenclature

Level	Name	Example(s)	Applies to…
Method (any combination of the different levels)			
1	Approach	RBC, NLI, FEM, LLM	*n/a*
Sub-method (any combination of a *setup* and *dataset variant;* a method for a given *approach*) **Setup** (levels 2 and 3; any combination of a *model, query, task (type)*, and/or *prompt*)			
2	Model	BART-large-mnli, ClimateBERT, GPT-4o	NLI, FEM, LLM
	or Query	Baseline query, data-driven query	RBC
3	Task (type)	Labels (‘adaptation’, ‘mitigation’), task type (multi-class)	NLI
	or Prompt	P1 (concise), P2 (specific)	LLM
Dataset variant (levels 4 and 5; any combination of a *dataset* and *data strategy*)			
4	Dataset	Dataset 1 (full chunks), dataset 2b (sub-chunks, resolved)	*all*
5	Data strategy	CST0, CST70, filtered, unfiltered	*all*