www.nature.com/scientificdata

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Measuring China's Policy Stringency on Climate Change for 1954–2022

Bo Li[1,6], Enxian Fu[2,6], Shuhao Yang[3], Jiaying Lin[4], Wei Zhang[1]✉, Jian Zhang[5]✉, Yaling Lu[1], Jiantong Wang[1] & Hongqiang Jiang[1]

Efforts on climate change have demonstrated tangible impacts through various actions and policies. However, a significant knowledge gap remains: comparing the stringency of climate change policies over time or across jurisdictions is challenging due to ambiguous definitions, the lack of a unified assessment framework, complex causal effects, and the difficulty in achieving effective measurement. Furthermore, China's climate governance is expected to address multiple objectives by integrating main effects and side effects, to achieve synergies that encompass environmental, economic, and social impacts. This paper employs an integrated framework comprising lexicon, text analysis, machine learning, and large-language model applied to multi-source data to quantify China's policy stringency on climate change (PSCC) from 1954 to 2022. To achieve effective, robust, and explainable measurement, Chain-of-Thought and SHAP analysis are integrated into the framework. By framing the PSCC on varied sub-dimensions covering mitigation, adaptation, implementation, and spatial difference, this dataset maps the government's varied stringency on climate change and can be used as a robust variable to support a series of downstream causal analysis.

Background & Summary

Significant progress has been globally achieved in climate policy, as evidenced by extensive coverage, diversified mechanisms, and improved effectiveness in the field of climate governance[1]. Notably, those policies have demonstrated a discernible impact on climate change mitigation and adaption, particularly in specific countries, sectors, and technologies, including the noteworthy achievement of avoiding emissions equivalent to several GtCO$_2$-eq yr^{-1} [2,3]. However, considering the temperature control targets, there remains a substantial gap between the expected effects and the current policy supply[4].

To address the gap in achieving temperature control targets, systematic policy settings and packages are expected to balance synergies among climate, social, and economic systems simultaneously. In the near term, integrating climate change action with sustainable development and low-carbon transition, aligning climate change targets with multiple economic, environmental, and social objectives, and combining adaptation with mitigation options are the three major concerns[5–7]. Several researches has demonstrated that the effectiveness of policies lies in the right mix rather than in the quantity of policies[1,8]. Consequently, it is urgent to analyze the complex causal effects underlying these interactions to support the design and implementation of policy packages that take into account the local context and macro perspective.

However, a conceptual paradox emerges with the current definition of climate policy, which is often based on principles of empirical experience. Traditionally, climate policy encompasses policies formulated specifically to tackle climate change and can broadly be categorized into two main types: those designed to minimize greenhouse gas (GHG) emissions — climate change mitigation[9] — and those intended to minimize risks and seize new opportunities — climate change adaptation[10]. This traditional definition and taxonomy are too narrow to fully capture the breadth of human-driven efforts to address climate change. As a result, number of policies whose main objective is not climate change are inevitably neglected in current research. For example, when an economic policy addresses a pre-existing market failure that is made worse by climate change, it can inadvertently

[1]Chinese Academy of Environmental Planning, State Environmental Protection Key Laboratory of Environmental Planning and Policy Simulation, Beijing, 100043, China. [2]Harvard University, Harvard Kennedy School, Cambridge, 02138, USA. [3]Chinese Academy of Environmental Planning, Centre of Situation Analysis and Planning Assessment, Beijing, 100043, China. [4]World Resources Institute China, Sustainable Cities Program, Beijing, 100027, China. [5]School of Government, Central University of Finance and Economics, Beijing, 100081, China. [6]These authors contributed equally: Bo Li, Enxian Fu. ✉e-mail: zhangwei@caep.org.cn; zjcufe@cufe.edu.cn

support climate mitigation or adaptation by improving the system's overall resilience whether its primary purpose is climate change mitigation/adaption or not[4,11,12].

The major challenges above have become more concrete and urgent in China. On the one hand, China's historic coal consumption, greenhouse gas (GHG) emissions, and GDP growth have shown uptrend fluctuations since the first Five-Year Plan was initiated in 1953. During 1953–1970, the proportion of coal consumption exceeded 80%–95% during the energy-intensive industrialization period[13–16]. Before officially joining global climate governance around the 1990s, China had already accumulated considerable experience with various practices that can now be recognized as precursors to modern climate change mitigation and adaptation strategies. For example: (1) The 1957 Instructions on Strengthening the Leadership of Saving Coal Use advocated optimizing ignition technology and equipment improvements; (2) The 1955 Resolution on Comprehensive Planning for the Eradication of Water Disasters and Development of Water Conservancy in the Yellow River implemented agricultural and forestry soil improvement measures and water conservation efforts; (3) The 1958 Instructions on Large-scale Afforestation throughout the Country aimed to double the country's forest area within a decade. These measures, although not framed within the contemporary context of climate change, can be identified as unintended but effective mitigation and adaptation options in today's climate change mainstream discourse.

On the other hand, China has established a comprehensive top-down administrative system for climate governance under the strategic guidance of the Central Committee of the Communist Party of China (CCCPC) and the supervision of the National People's Congress of China (NPCC)[13,17,18]. This system operates under the unified leadership of the National Leading Group for Climate Change (NLGCC), with the central government managing policy implementation, and with the participation of all relevant departments and extensive engagement from localities and industries. Thus, efforts to tackle climate change have been scattered and overlapped across various policies and sectors[19–21]. Moreover, following global trends and domestic demands, China's climate change governance has developed into "Integrating Carbon Reduction, Pollution Reduction, Green Expansion, and Sustainable Growth" as outlined in the CCCPC's decision-making at the 20th CPC National Congress. Consequently, designing and implementing climate policy packages concerning multiple objectives expect robust empirical evidence locally while a single definition of climate policy is insufficient to capture the breadth of national actions and historical policy behaviors.

Research on climate policies and policymaking has expanded upon reviewing the literature in recent years. First, there is growing recognition that climate change action occurs in the context of multiple climate and development objectives[10]. Second, there is growing attention to enabling transitions over time[22]. However, limited research has been devoted to addressing the research gaps and their reification in China to promote further effective interventions: (1) Current datasets on climate policies are often initiated based on subjective criteria without a clear assessment of whether they comprehensively cover all effective policy measures. (2) Mitigation-oriented policies and adaptation-oriented policies are often studied separately in the literature, which leads to a lack of robust data instruments to understand the co-benefits and trade-offs between these two mainstreams. This gap hinders the ability to assess how these policies can either accelerate or offset climate change actions. Integrating the analysis of mitigation and adaptation policies is crucial for developing comprehensive strategies. (3) In the context of China's climate governance, there is a lack of deeper understanding of how policies are implemented and diffused, especially when taking into account spatial differences. Within the same policy, the intensity of policy objectives and measures can vary, and this variation is more pronounced in terms of space. For instance, some policy provisions are more focused on certain regions, which is particularly prominent in adaptation policies. (4) There is a lack of detailed decomposition for existing options from global mainstreams, such as those outlined in AR6, to match with local climate governance frameworks. For example, afforestation, which is a major area in China's climate governance handled by the National Forestry and Grassland Administration (NFGA), is often integrated with ecosystem restoration corresponding to the Ministry of Ecology and Environment (MEE) into a single option in AR6. This integration can obscure the distinct benefits and challenges associated with each activity, resulting in evaluation bias.

Overall, as climate policies gain increasing prominence, we argue that the theoretical understanding of climate policy needs to expand in three key dimensions. First, the benchmark for identifying climate policies should broaden to encompass all human interventions and actions that have both direct and indirect potential effects, not just proactive measures. Second, policy should be meticulously decomposed to precisely identify their main and side effects, accounting for both intended and unintended consequences. Third, historical evidence supports this broader perspective: global CO_2 emissions have shown observable fluctuations since 1850[23,24], with many of these changes driven by policy interventions. This historical pattern provides valuable context for developing future climate governance frameworks, even though formal global climate governance only emerged in the 1980s.

Therefore, in this paper, we define "Policy Stringency on Climate Change" (PSCC) to depict the stringency of those policies with the positive effects on climate change, regardless of whether their main targets are climate change or not. This expanded definition enhances the theoretical connotation of climate policy from the perspective of objective posterior, enabling measurement, causal inference, and the development of comprehensive policy packages. Based on this clear and specific definition of PSCC, we utilized a framework integrating feature lexicons, text analysis, machine learning, and large language models to construct a dataset of China's PSCC for the period 1954–2022. This dataset includes 2,216 policies selected from over 24,000 candidates which include governmental publication archives, third-party databases, and academic datasets. The resulting dataset is the most comprehensive to date for China at the national level, offering insights into China's hierarchical climate governance structure and governmental behavior concerning climate change. This dataset can be parallelly applied to other countries and vertically scaled to provincial and city levels for broader applicability.

Our research stands out for several innovative contributions: First, we expand the definition of climate policy to address the ambiguity that can lead to potential biases in causal effects. We then develop an integrated

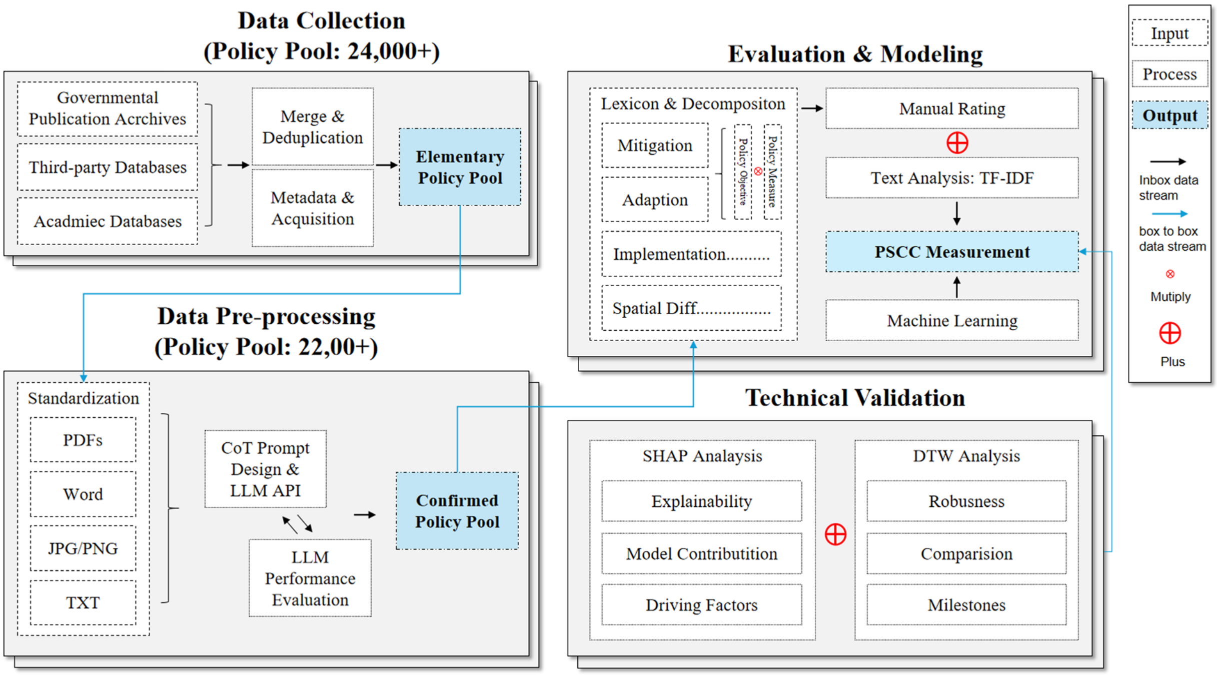


Fig. 1 Schematic. The framing of this study is composed of four parts Data collection, Data Pro-processing, Evaluation and Modeling, and Technical Validation.

framework that combines lexicon-based methods, text analysis, machine learning, and large language models to measure policy stringency on a policy-by-policy basis. Second, we introduce and integrate a Large Language Model (LLM) into our framework. Climate governance encompasses a broad spectrum of fields, making some policies with unintended positive effects on climate change neglected. To address the challenge of identifying potential target policies, we leverage the world comprehension and generalization abilities of LLMs through systematically designed prompt engineering. Techniques such as Chain-of-Thought (CoT) and integrated evaluation metrics are employed to guide the LLM in reading, understanding, and further inferring which policies should be selected for our subsequent analysis accurately. Third, we develop a robust measurement variable. Our Policy Stringency on Climate Change (PSCC) for each policy can be broken down into various sub-dimensions, providing robust variables for comprehensive causal effects analysis. Fourth, we enhance the explainability of our technical validation by introducing SHapley Additive exPlanations (SHAP) analysis and Dynamic Time Warping (DTW) into our framework to provide detailed insights into the model's decision-making process and ensure that our results are transparent and interpretable. These innovations collectively advance the field of climate policy analysis, offering a more powerful approach to evaluating policy stringency and effectiveness.

Methods

Our framework is constructed with four modules: Data collection, Data pre-processing, Evaluation and modeling, and Validation. And it's presented in Fig. 1.

Data collection.

Potential policies identification. As previously concluded, climate action involves multiple sectors, organizations, and authorities[22]. Adhering to the definition of Policy Stringency on Climate Change (PSCC), our goal was to incorporate all policies that could potentially impact climate change into our research boundary. To get potential policies enrolled comprehensively across various climate-related domains, without regard to the primary objective or effect, we adopted an inclusive approach, initially including a wide array of policies regardless of their categorization or domain, provided they were issued at the national or central level. Following a detailed investigation, we identified and confirmed 8 data sources, which include governmental publication archives, third-party databases, and academic datasets; these sources and their basic statistics are detailed in Table 1.

Metadata database and data acquisition. Based on the initial dataset obtained from the previous step, our next tasks involve building a metadata database and developing a data acquisition process. The first step is to establish a metadata database that captures the essential identification information of each policy, such as a unique identification (UID), title, category, and other relevant details, which can be found in Table 2. This database serves as a centralized repository for organizing and managing the metadata associated with climate policies. The second step involves data acquisition. To achieve this, we developed a task-oriented web scraper using the RSelenium package created by Harrison *et al.*[25]. RSelenium is a powerful tool that leverages Selenium, a widely recognized framework for automating web browsers, enabling us to collect detailed policy information from various online sources efficiently. To ensure that the metadata database effectively maps to the policy documents gathered from multiple sources, each policy document is assigned a unique identifier (UID) upon ingestion into the system,

Dataset	Description	Records	Selected Records	Format	Reference
China's State Council Gazette Archive System (CSCGAS)	CSCGAS, published by the General Office of the State Council, is a domestic and international government publication that accurately releases key documents including administrative regulations, decisions, and orders, along with important rules and files from its various departments, and personnel changes.	Started since 1954, with each year comprising dozens of issues, and each issue containing varying numbers of policies.	All selected from 1954–2022.	For data prior to 2000, the format is PDF files containing images. For data post-2000, the format is embedded HTML within web pages.	https://www.gov.cn/gongbao
The State Council Policy Document Library (SCPDL)	SCPDL contains various policy documents issued by the State Council with China's policies, regulations, and guidelines across diverse areas.	28,720	All selected from 1954–2022, 25,624.	Embedded HTML within web pages.	https://sousuo.www.gov.cn/zcwjk/policyDocumentLibrary?q=&t=zhengcelibrary&orpro=
Chinese Government Bulletins and Journal Literature Database (CGBJLD)	CDCGBJL includes 223 types of government bulletins, serving as an authoritative and standardized one-stop search platform for official government files.	587,140	Only central and national level policies are included in our pool, 23,456.	The dataset consists of two types of PDF files. The first type is primarily text-based PDFs, while the second type is image-based PDFs, including scanned images of Traditional Chinese text for some early-year data.	https://zhengbao.cnki.net/index?sysid=23
PKULAW	The Peking University Legal Information Retrieval Platform, known as PKULAW, is a comprehensive legal information search platform in China.	Over 1.47 million regulatory documents from central and local governments	Only central and national level records are included in our pool.	Text-based PDFs.	https://www.pkulaw.com/advanced/law/chl
Global Climate Change Mitigation Policy Database (GCCMPD)	The GCCMPD is a global climate change mitigation policy dataset covering a large range of policies from 216 entities.	73,625	Only policies belonging to China are included, 1,640.	Multi-format including MySQL, MongoDB, and EXCEL.	https://doi.org/10.6084/m9.figshare.22590028.v2
China's Low-Carbon Policy Intensity Dataset (CLCPID)	The CLCPID is derived from a specialized low-carbon policy inventory, covering China's manufacturing industries from the national to the prefectural levels between 2007 and 2022.	7,282	Only central and national level records are included in our pool, 91.	Multi-format including STATA and EXCEL.	https://doi.org/10.6084/m9.figshare.c.6761106.v1
China's Environmental Policy Intensity Dataset (CEPID)	The CEPID is derived from China's national environmental policy dataset between 1978 and 2019.	1,912	All selected, 1,912.	EXCEL.	https://doi.org/10.6084/m9.figshare.16740376.v1
Memorabilia of the Communist Party of China (MCPC)	MCPC is an applied style for party and government organs, enterprises, institutions, and social organizations to record their important work activities or major events in their jurisdictions.	Annually since CPC established	Records from 1964 to 1979 selected as supplement, 16.	Embedded HTML within web pages.	http://cpc.people.com.cn/GB/64162/64164/4416064.html

Table 1. Data sources and basic information.

acting as a primary key that links the metadata record to the corresponding policy document. We also maintain a record of the source from which each policy document was obtained, which helps in verifying the authenticity and provenance of the documents and aids in resolving any discrepancies or inconsistencies. For each policy document, we store the URL where it was originally published, allowing users to access the full text directly and ensuring that the metadata is linked to the source material.

After completing the steps of potential policy identification, metadata database framing, and data acquisition, we have established two foundational databases. The first database records the identification information of all collected policies, while the second is a comprehensive policy pool with specific policy text. This pool is characterized by its multi-source, multi-format, multiple-theme, and multi-time scale nature. Our elementary policy pool contains over 24,000 records, with 9.3% sourced from governmental data, 90% from third-party data sources, and 0.7% from academic datasets.

Data pre-processing. *Multi-source data standardization.* In the realm of data pre-processing, multi-format data, including PDF, TXT, DOC, and JPG files, need to be standardized. Each format presents unique challenges that require specific handling to ensure uniformity and accessibility across the dataset. For standard PDF files

Field	Description
UID	A Unique Identifier (UID) is employed to distinguish individual policies and establish a mapping between the metadata database and specific policies. The UID is composed of two segments: the first part indicates the year of archiving, and the second part denotes a sequential number. For example, the UID "201317437" signifies that the policy was archived in the year 2013 and is the 17,437th entry in the overall sequence.
URL	A URL is the web address that points to the exact location of a policy document online. It is used to direct users to the specific web page where they can access the full text of the policy document.
Year	This refers to the calendar year during which a policy is officially enacted or promulgated by the governing body.
Policy Title	This denotes the formal name given to a policy document that succinctly describes its main focus or subject matter. The title serves as an identifier and provides insight into the scope and intent of the policy without having to read the entire document.
Policy Title in Chinese	Given the target audience is English-speaking but the subject matter is in Chinese, this term would refer to the exact translation of the policy's title from Chinese to English.
Category-I	From a policy science perspective, categorizing policies can be done based on the regulatory approach they adopt. Here we propose three categories: Command-and-Control Instrument (CCI), Market-based and mixed Instrument (MXI), and Voluntary Instrument (VOI).
Category-II	This classification is based on a lexicon system designed to match the most potentially relevant sub-dimensions with policy content. It serves as an introductory guide for further quantitative analysis in a specific field. Such as policy Implementation Plan of China Green Lighting Project could be categorized as Mitigation_Buildings_Efficient Lighting, Appliances And Equipment with unique code 132.

Table 2. Metadata Database Information.

containing primarily text, we use PDF parsing libraries to extract the text content. When dealing with PDFs that include images or scanned pages, we employ Optical Character Recognition (OCR) technology to help extract text from the images within the PDFs. For PDFs that contain traditional Chinese text, we utilize machine translation services to convert traditional Chinese into simplified Chinese. Once translated, we use the same PDF parsing tools mentioned above. For DOC files, we use document processing libraries to extract the text content. Any embedded images or tables are ignored or handled separately if necessary. Images containing text, such as JPG files, require OCR processing similar to PDFs consisting of pictures. By following these steps, we transform the files into uniform TXT files with a standardized naming convention. After extraction, we use regular expressions (Regex) in R to clean the text, removing residual noise and standardizing the output. This process ensures that the text is free of extra spaces, symbols, English numbers, and other noisy elements, and is rearranged into a tidy, standardized format where sentences are presented sequentially without unnecessary spacing or symbols. Examples of multi-format data standardization can be found in Table S1 (see Supplementary Information document).

LLM-based filtering and classifying system. Recent advancements in artificial intelligence have seen the rise of large language models (LLMs), which are emerging as a powerful tool for natural language processing (NLP). These models are capable of transforming the landscape of text and language data analysis by leveraging deep learning techniques, vast computational resources, and extensive training datasets to produce coherent and contextually relevant text[26]. In this way, many researchers have utilized LLMs in a variety of NLP tasks like question answering, natural language inference, named entity recognition, relation extraction, information extraction, and text classification[27]. One notable area of success is text classification, where LLMs have achieved remarkably high accuracy. In particular, their application in policy classification has demonstrated the models' superior ability to predict outcomes through prompt-based learning[28]. Unlike traditional supervised learning methods, which often require large, labeled datasets that can be expensive and time-consuming to create, LLMs utilize prompt learning as a more efficient approach. Traditional models typically aim to optimize the conditional probability $P(y|x;\theta)$, where y is the predicted output, x is the input text, and θ are the model's parameters[28–30]. LLMs use prompt learning as more efficient approach by modeling the probability $P(x;\theta)$ of input text, allowing the model to predict labels with minimal supervised data[30]. This shift in approach is encapsulated by the equation:

$$\hat{y} = \underset{y \in Y}{\mathrm{argmax}} P(y|x;\ \theta) \tag{1}$$

Where $\hat{y}$ is the predicted label, $P(y|x;\theta)$ is the conditional probability of output y, and Y is the set of possible labels.

Therefore, we utilized the advantage of LLMs to establish our LLMs-based filter and policy classification system. There are three major groups of LLM models: (1) small, encoder-only models like BERT that specialize in encoding semantic information[29]; (2) medium to large, decoder-only and encoder-decoder models like Text-to-Text Transfer Transformer (T5)[31], Grouped-Query Attention (GQA)[32], Llama models[33], etc., which are capable of autoregressive text generation; (3) and large proprietary, decoder-only models like GPT-4o[34] and Qwen-72B[35], which can only be run using third-party APIs. Recent comparisons demonstrate that decoder-only models achieve superior performance across numerous benchmarks. These models are increasingly favored for their streamlined pre-training process and more efficient model size[36]. Therefore, we chose to use GPT-4o as the primary model and Qwen-72B as a prompt augmentation model. Both models can be accessed via third-party APIs to help us filter and classify policies. Our workflow is presented in the Fig. 2.

The first task involves filtering policies based on their potential to mitigate or adapt to climate change. Large language models (LLMs) demonstrate exceptional ability to identify such policies, even in the absence of

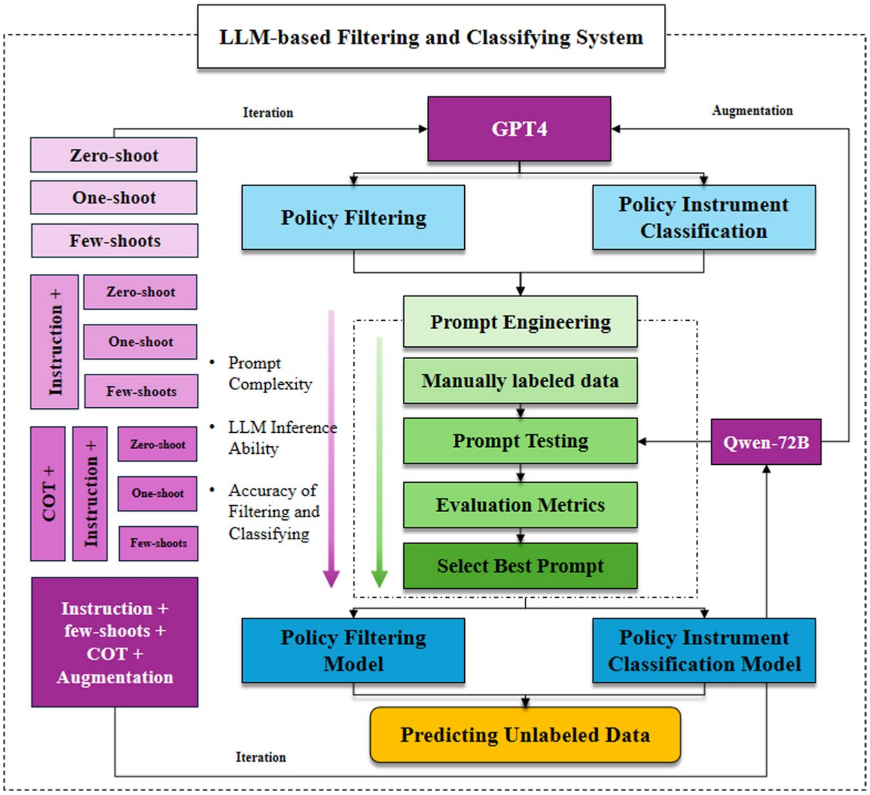


Fig. 2 LLM-based Filtering and Classifying System. This diagram outlines the LLM-based filtering and classification system used to process and categorize policies. The system is built around GPT-4o for primary processing, with augmentation support from Qwen-72B. The flow begins with Policy Filtering and Policy Instrument Classification tasks. The core of the system is the Prompt Engineering stage, which iteratively refines prompts using zero-shot, one-shot, and few-shot learning techniques. Within prompt engineering, different levels of prompt complexity, including Instruction, Chain-of-Thought (COT), and Augmentation strategies, are tested. Manually labeled data serves as the foundation for Prompt Testing and evaluation through Evaluation Metrics like accuracy and F1-score, allowing for the selection of the optimal prompt. The finalized models for Policy Filtering and Policy Instrument Classification are then applied to unlabeled data for classification predictions. The prompt engineering and testing loop iterates until achieving the best-performing prompt configuration, ensuring reliable filtering and classification results.

keywords like 'climate change' or 'carbon emissions. This is achieved through systematic prompt design and the integration of cutting-edge LLM technologies such as Chain-of-Thought[37]. This capability is crucial for capturing a comprehensive and nuanced understanding of policy impacts. By leveraging advanced natural language processing, LLMs can detect subtle references and contextual cues that traditional keyword-based methods might miss, ensuring a more thorough and accurate assessment of policy relevance.

The second task involves classifying policies based on their policy instruments: command-and-control instruments (CCI), market-based mixed instruments (MXI), and voluntary instruments (VOI). CCI corresponds to laws and regulations, MXI to instruments like funding, taxes, and international cooperation, and VOI to indirect measures such as public awareness campaigns and educational initiatives[38–44]. The key objective of this task is to accurately classify policies into these categories and make reliable inferences about unlabeled objects. Traditional keyword searches often fall short in capturing the nuanced characteristics of each policy instrument, making large language models (LLMs) particularly well-suited for this task. LLMs can identify and understand the latent features of each policy instrument, ensuring a more comprehensive and accurate classification.

For both the filtering and classification tasks, we employed a comprehensive approach leveraging prompt learning, in-context learning, instruction learning, and chain-of-thought reasoning[37,45,46]. We conducted rigorous prompt engineering, starting from zero-shot learning, progressing through instruction learning, COT reasoning, and few-shot learning with augmentation[29]. For the filtering task, we focused on identifying policies with the potential to mitigate or adapt to climate change. We trained two models on 800 policies manually labeled by multiple experts.Among the 800 policies, 400 labeled as 1 were selected based on policy instruments and policy dimensions. Similarly, 400 policies labeled as 0 were chosen based on two perspectives to ensure the repetitiveness: (1) policies unlikely to have a significant environmental impact, and (2) policies frequently misclassified by LLMs. Additionally, we introduced Qwen as a prompt augmentation model to help select representative samples and refine the prompts. After iterating on the prompt engineering process, we selected the best-performing prompt and applied it to filter the remaining unlabeled policies (Table S4 in Supplementary Information document details the best prompt). For the classification task, we applied the same strategy to categorize policies by their instrument types: command-and-control instruments (CCI), market-based mixed instruments (MXI), and voluntary instruments (VOI). We applied our system to policies filtered in the first task (Table S5 in Supplementary Information document shows the usage of the best prompt).

LLM performance assessment. The evaluation of large language models (LLMs) typically follows two common approaches: automatic evaluation and human evaluation, which are often combined in research for comprehensive model assessments[47]. Automatic evaluation, the more frequently used method, relies on computational metrics and tools to assess the model's performance without human intervention. This approach is favored for its efficiency, consistency, and ability to reduce human bias. Common metrics used include accuracy-based metrics, such as Exact Match (EM), Quasi-exact match, F1-score, and ROUGE score[48], which are extensively applied across various NLP tasks. F1-score and its components, precision, and recall, are particularly useful for both binary and multi-class classification tasks[28,29,49]. In contrast, EM and ROUGE are more suited for tasks like text generation, summarization, and machine translation.

Notably, Dong *et al.*[50] introduced LLMs to assess China's low-carbon policy intensity, using accuracy as a primary evaluation metric. However, relying solely on accuracy can be misleading, particularly in imbalanced datasets, where dominant majority classes may overshadow the model's performance on minority classes[51]. To address this, we incorporated a combination of accuracy, precision, recall, and F1-score to achieve a more precise evaluation. This suite of metrics provides a comprehensive assessment of prompt-specific model performance and the model's ability to accurately filter policies based on climate impact and classify specific policy instruments.

The metrics were calculated using both binary-average for binary classification tasks and weighted-average for multi-class classification tasks, where class imbalance needed to be considered[51]. In our system, for policy filtering, binary-average metrics were used to distinguish between policies with and without climate change effects. For the classification of policy instruments (command-and-control, market-based, and voluntary instruments), weighted-average metrics were employed. This combination of metrics ensures that the model's performance is well-rounded and robust against class imbalances. Details on the metric formulas are provided in Note S1 (see Supplementary Information document).

Tidying a document-term matrix. In previous steps, we obtained a metadata database mapping the policy pool and a policy pool after Large Language Model (LLM) filtering and classification, which will serve as our research object for conducting NLP algorithms with text analysis. With the policies prepared in the secondary policy pool, we are going to convert them into a document-term matrix (DTM) format to facilitate a series of downstream tasks. The document-term matrix (DTM) is a common structure used in text mining packages, where each row represents one document (such as a policy document), each column represents one term, and each value typically contains the number of appearances of that term in that document. Converting our policies into this format allows us to rely on both existing text mining packages and the suite of tidy tools to perform our analysis. Such as the tidy() function transforms a document-term matrix into a tidy data frame, while cast() converts a tidy one-term-per-row data frame back into a matrix, thus enabling a series of text analysis, mining, and visualization tasks. These packages are very useful in text mining applications, and many existing text datasets are structured according to these formats.

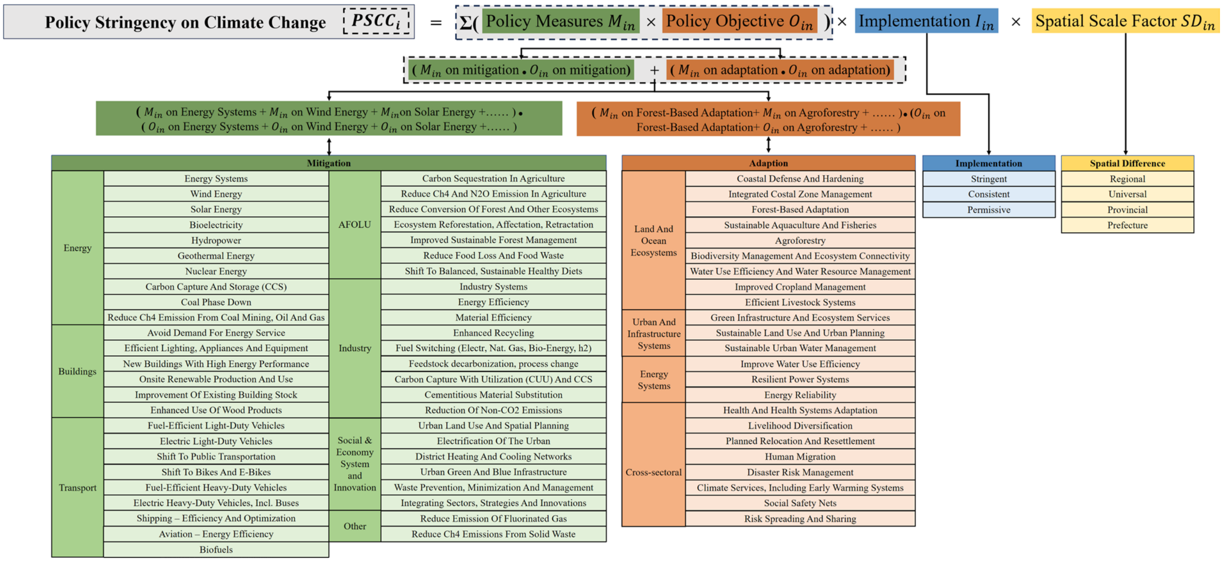
PSCC evaluation and modeling. To address limitations inherent in simple word frequency analysis, we developed an integrated framework comprising three core components: manual rating, policy-by-policy text analysis, and machine learning modeling to evaluate Policy Stringency on Climate Change (PSCC). This approach builds on the methodology of Zhang *et al.*[52,53], with an expanded formula for calculating PSCC as follows:

$$PSCC_i = \sum (M_{in} * O_{in}) * I_{in} * SD_{in} \tag{2}$$

In this formula, i represents the specific policy, and n denotes the number of sub-dimensions within mitigation and adaptation. The term M_{in} indicates the aggregated stringency rating of policy measures within each sub-dimension of mitigation and adaptation. Similarly, O_{in} represents the aggregated stringency rating of policy objectives within these same sub-dimensions. Meanwhile, The aggregated implementation score I_{in} represents the degree of enforcement or rigor associated with implementing these measures across all relevant sub-dimensions. Additionally, SD_{in} represents the spatial scale factor, which adjusts the overall stringency score based on the spatial difference and geographical specificity mentioned in the policy.

We define the spatial scale factor as an indicator that reflects geographical specificity in policies. Policies that focus on distinct areas such as the Yangtze River region, specific provinces like Sichuan, or individual cities like Beijing receive higher spatial scale factor values. It is derived through Principal Component Analysis (PCA) applied to TF-IDF values across four spatial categories: universal, regional, provincial, and prefecture. This factor serves to adjust the manual rating according to the policy's geographic scope, where higher values indicate more localized, targeted, or geographically specific policies. The detailed calculation of spatial scale factor using PCA is putting Note S2 (see Supplementary Information document).

To clarify the calculation, we illustrate by graph on the Fig. 3. We also provide a manual-based benchmark example: a policy has (1) a stringency rating of 3 in policy measures of mitigation and a rating of 5 in policy objective of mitigation, (2) a stringency rating of 5 in policy measures of adaption and policy objective of mitigation, (3) a stringent implementation score of 5, (4) if the spatial scale factor is 1.24. The overall manual rating is thus calculated: $((3 \times 5) + (5 \times 5)) \times 5 \times 1.24 = 99.2$.


Policy Stringency on Climate Change $PSCC_i$ = Σ(Policy Measures M_{in} × Policy Objective O_{in}) × Implementation I_{in} × Spatial Scale Factor SD_{in}
(M_{in} on mitigation • O_{in} on mitigation) + (M_{in} on adaptation • O_{in} on adaptation)
(M_{in} on Energy Systems + M_{in} on Wind Energy + M_{in}on Solar Energy +......) • (O_{in} on Energy Systems + O_{in} on Wind Energy + O_{in} on Solar Energy +......)
(M_{in} on Forest-Based Adaptation+ M_{in} on Agroforestry + ...)• (O_{in} on Forest-Based Adaptation+ O_{in} on Agroforestry + ...)

Mitigation			
Energy	Energy Systems	AFOLU	Carbon Sequestration In Agriculture
	Wind Energy		Reduce Ch4 And N2O Emission In Agriculture
	Solar Energy		Reduce Conversion Of Forest And Other Ecosystems
	Bioelectricity		Ecosystem Reforestation, Affectation, Retractation
	Hydropower		Improved Sustainable Forest Management
	Geothermal Energy		Reduce Food Loss And Food Waste
	Nuclear Energy		Shift To Balanced, Sustainable Healthy Diets
	Carbon Capture And Storage (CCS)	Industry	Industry Systems
	Coal Phase Down		Energy Efficiency
	Reduce Ch4 Emission From Coal Mining, Oil And Gas		Material Efficiency
Buildings	Avoid Demand For Energy Service		Enhanced Recycling
	Efficient Lighting, Appliances And Equipment		Fuel Switching (Electr, Nat. Gas, Bio-Energy, h2)
	New Buildings With High Energy Performance		Feedstock decarbonization, process change
	Onsite Renewable Production And Use		Carbon Capture With Utilization (CUU) And CCS
	Improvement Of Existing Building Stock		Cementitious Material Substitution
	Enhanced Use Of Wood Products		Reduction Of Non-CO2 Emissions
Transport	Fuel-Efficient Light-Duty Vehicles	Social & Economy System and Innovation	Urban Land Use And Spatial Planning
	Electric Light-Duty Vehicles		Electrification Of The Urban
	Shift To Public Transportation		District Heating And Cooling Networks
	Shift To Bikes And E-Bikes		Urban Green And Blue Infrastructure
	Fuel-Efficient Heavy-Duty Vehicles		Waste Prevention, Minimization And Management
	Electric Heavy-Duty Vehicles, Incl. Buses		Integrating Sectors, Strategies And Innovations
	Shipping – Efficiency And Optimization	Other	Reduce Emission Of Fluorinated Gas
	Aviation – Energy Efficiency		Reduce Ch4 Emissions From Solid Waste
	Biofuels		

Adaption	
Land And Ocean Ecosystems	Coastal Defense And Hardening
	Integrated Costal Zone Management
	Forest-Based Adaptation
	Sustainable Aquaculture And Fisheries
	Agroforestry
	Biodiversity Management And Ecosystem Connectivity
	Water Use Efficiency And Water Resource Management
	Improved Cropland Management
	Efficient Livestock Systems
Urban And Infrastructure Systems	Green Infrastructure And Ecosystem Services
	Sustainable Land Use And Urban Planning
	Sustainable Urban Water Management
Energy Systems	Improve Water Use Efficiency
	Resilient Power Systems
	Energy Reliability
Cross-sectoral	Health And Health Systems Adaptation
	Livelihood Diversification
	Planned Relocation And Resettlement
	Human Migration
	Disaster Risk Management
	Climate Services, Including Early Warming Systems
	Social Safety Nets
	Risk Spreading And Sharing

Implementation
Stringent
Consistent
Permissive

Spatial Difference
Regional
Universal
Provincial
Prefecture

Fig. 3 Policy Stringency on Climate Change (PSCC) Quantification Framework. This figure illustrates the expanded structure used for quantifying the Policy Stringency on Climate Change (PSCC) score based on four main pillars: Policy Measures, Policy Objectives, Implementation, and Spatial Scale Factor. Following the IPCC AR6 framework, mitigation (green) and adaptation (orange) dimensions are each broken down into specific sub-dimensions (e.g., Energy Systems, Land and Ocean Ecosystems). Each sub-dimension within mitigation and adaptation is evaluated based on the rigor of policy measures and policy objectives. Implementation (blue) is assessed separately across three levels: Stringent, Consistent, and Permissive. Spatial Scale Factor (yellow) incorporates geographic categories (Regional, Universal, Provincial, Prefecture) and is derived using Principal Component Analysis (PCA) to adjust PSCC based on policy impact breadth.

Multi-layered, cross-sectoral lexicon build-up. Based on the IPCC's AR6 reports[9] and comprehensive literature reviews that details in Table S3 (see Supplementary Information document), we developed a multi-layered, cross-sectoral lexicon structured around four Tier-I dimensions: Mitigation, Adaptation, Implementation, and Spatial Difference. Each of these dimensions is further subdivided into the following Tier-II sub-dimensions: (1) Mitigation (Energy, AFOLU, Buildings, Transport, Industry, Urban systems, and Others); (2) Adaption (Land and ocean ecosystems, Urban and infrastructure systems, Energy systems, and Cross-sectoral); (3) Implementation (Consistent, Permissive, and Stringent); (4) Spatial Difference (Universal, Regional, Provincial, prefecture). Furthermore, each Tier-II sub-dimension is broken down into 50 and 23 Tier-III sub-dimensions for Mitigation and Adaptation, respectively. Moreover, each Tier-III sub-dimension is then further divided into "Policy objectives" and "Policy measures". Finally, we obtained 4 Tier-I, 18 Tier-II, 73 Tier-III, and 146 Tier-IV sub-dimensions in total. The framework for the lexicon is detailed in Fig. 3. Following the lexicon preparation framework, we invite researchers from both environmental science and social science to build up the lexicon separately and obtain a final version after merging and cross-checking.

Manual rating. For the manual-based benchmark, each policy is evaluated across four core aspects: policy measures, policy objectives, implementation, and spatial differences. Policy measures and objectives are assessed within both mitigation and adaptation sub-dimensions, while implementation is evaluated across three levels: stringent, consistent, and permissive. The manual ratings for policy measures, policy objectives, and implementation utilize a stringency scale from 1 to 5. Spatial differences are incorporated through Principal Component Analysis (PCA) applied to the TF-IDF values, with further details provided in the following section. We opted for the 1-3-5 scale rather than a linear 1-2-3 scale because it offers clearer differentiation between levels of policy stringency, reducing ambiguity and encouraging more decisive judgments from evaluators. Research suggests that scales with wider numerical gaps help minimize the tendency for respondents to choose neutral or middle points, thus providing sharper distinctions in intensity or importance[54,55]. This approach helps evaluators avoid 'central tendency bias,' where ratings gravitate toward the middle, potentially masking meaningful differences in policy impact. The 1-3-5 scale also enhances the identification of policies with significant or minimal impact by making middle ratings less frequent. This method is particularly effective in decision-making contexts where clear prioritization is necessary[56]. As outlined by Zhang *et al.*[53], Table S6 (see Supplementary Information document) provides scoring criteria for policy measures in mitigation and adaptation, while Table S7 (see Supplementary Information document) describes criteria for policy objectives n mitigation and adaptation.

Corresponding to the rating of implementation particularly, we rate it separately as a key component. This approach aligns with the view that successful climate policies require integrated governance, which includes planning, monitoring, and evaluation stages[57–60]. Table S8 (see Supplementary Information document) outlines criteria for evaluating policy implementation effectiveness, categorizing implementation as stringent, consistent, or permissive based on specific guiding language. The research underscores the significant impact that policy language and implementation levels have on policy effectiveness[61,62].

Corresponding to the spatial difference behind those policies' intentions, we introduce a spatial scale factor that accounts for geographic specificity. This factor is divided into four levels: universal, regional, provincial, and prefecture. Due to the complexity of manually evaluating spatial distinctions across these levels — for instance, when a document contains language that spans multiple levels — it is challenging to assign a single, accurate rating. To incorporate this spatial scale factor into our manual scoring process, we apply Principal Component Analysis (PCA) on the TF-IDF values derived from spatial keywords in our lexicon for each level. PCA reduces the dimensionality from four to three. And we aggregate these values into a spatial scale factor. The final policy stringency score is then calculated by multiplying this spatial scale factor with the manual ratings, ensuring that the score reflects the policy's spatial specificity.

To ensure rigorous evaluation, we assembled and trained a team to independently rate each policy across these criteria. Multiple raters assessed each policy to maintain high interrater reliability. These ratings serve a dual purpose: they capture the emphasis a policy places on specific measures and objectives while addressing challenges associated with weighting decisions in indicator construction.

Text analysis. Word frequency, a widely used technique in text analysis, has demonstrated the word frequency effect, indicating that high-frequency words are processed more efficiently[63]. The utilization of text-analysis approaches for constructing indicators has garnered increasing attention in recent literature[64–66]. The presence of long-tailed distributions in natural language corpora has led to the exploration of the relationship between word usage frequency and rank where invisible topics are revealed, known as Zipf's law[67]. However, the dependency of frequency counts on corpus size and the potential noise introduced by stop-words necessitate caution in word frequency analysis. To provide high-quality machine learning input, TF-IDF has been proven effective in characterizing the themes, topics, and attention within the text[68,69]. There are several key preprocessing steps to enhance TF-IDF analysis. (1) Stemming or lemmatization unifies word variants, focusing on essential vocabulary. (2) Removing punctuation and special characters, coupled with text normalization, ensures analysis precision by standardizing formats and excluding non-relevant elements. (3) Addressing missing data enriches insights into policy stringency against climate change. This streamlined approach underpins the TF-IDF algorithm's effectiveness in identifying critical terms for informed decision-making. The formula of the TF-IDF algorithm is detailed as follows:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

$$IDF_i = \ln\left(\frac{n\ documents}{n\ documents\ containing\ term}\right) \tag{4}$$

$$TF_{i,j} \times IDF_i = \frac{n_{i,j}}{\sum_k} \times \ln\frac{|D|}{\{|j:\ t_i \in d_i|\}} \tag{5}$$

where $TF_{i,j}$ denotes the number of occurrences of i in j, IDF_i denotes the number of documents containing i, D denotes total number of documents.

Machine learning model. After having the TF-IDF value derived from the lexicon, we adopt from the past intensive experience of text analysis approach with machine learning tools to facilitate the construction and validation of our PSCC[52,70]. Given the presence of complex nature of our lexicon, both traditional algorithms and machine learning algorithms can be deployed. In this study, we choose both traditional algorithms and machine learning algorithms: linear regression (LM), Elasticnet (Elastic), Lasso (Lasso), support vector machine (SVM), eXtreme gradient boosting tree (XgbTree), cubist regression model (Cubist), a bagging wrapper for multivariate adaptive regression splines via the earth function (BagEarth), and random forest (RF).

Linear regression (LM) serves as the baseline for comparing algorithmic performance. Elasticnet (Elastic) and Lasso (Lasso), with their feature selection capabilities, automatically focus on the most predictive features, combining Lasso's and Ridge's traits[71]. To adeptly manage high-dimensional data, we employ support vector machine (SVM) and models including eXtreme gradient boosting with tree booster (XgbTree), random forest (RF), and cubist regression model (Cubist). Additionally, the ensemble model BagEarth, a bagging approach for multivariate adaptive regression splines, addresses non-linear relationships in complex data[72]. Traditional and machine learning algorithms were both trained on the training dataset, comprising 70% of the data, and evaluated on the test dataset, which constituted the remaining 30%. To estimate the error more accurately, we implemented 10-fold cross-validation, repeated three times. Hyper-parameter optimization was conducted during the model training phase for the SVM, XgbTree, Cubist, Random Forest, and Bagearth models.

We opt for Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) as metrics to assess model performance using the test dataset in alignment with continuous predictive outcomes. The formulations for MAE and RMSE are provided below:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{6}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right| \tag{7}$$

Here, y_i denotes the observed value for the i-th instance, $\overline{y}_i$ represents the mean of the observed values, and $\hat{y}_i$ is the predicted value for the i-th instance. We employ both RMSE and MAE because, while MAE is indifferent to the error direction, RMSE is more sensitive to outliers. The combination of these metrics provides a comprehensive standard for robust performance evaluation.

To give a more interpretative explanation of our predictive model. We considered the SHapley additive exPlanations (SHAP), which is introduced by Lundberg and Lee[73]. The SHAP is rooted in cooperative game theory and was originally formulated by Shapley[74]. It provides an equitable framework for distributing payoffs among players, thereby serving as an excellent tool for discerning the influence of individual features on the final prediction of a model. Compared to Permutation Feature Importance used in other literature[52,75–77], the SHAP values are mainly based on the magnitude of feature attributions instead of the decrease in model performance. Therefore, it is the only solution that satisfies Efficiency, Symmetry, Dummy, and Additivity[73]. The SHAP explanation model (g) is articulated as follows:

$$g(z) = \Phi_0 + \sum_{i=1}^{M}\Phi_i z_i \tag{8}$$

Here, $z \in \{0,1\}^M$ with M denoting the total number of features. The term Φ_0 is the base prediction when no features are present. z_i is a binary indicator for the presence (1) or absence (0) of the i-th feature in a given instance. The expected prediction $f_x(S) = E[f(x)|x_s]$ over a subset of features S, allows the calculation of the SHAP value Φ_i as follows:

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f_x(S \cup \{i\}) - f_x(S)] \tag{9}$$

In this expression, S represents a subset of the full feature set F, excluding the i-th feature, while $f_x(S \cup \{i\})$ and $f_x(S)$ reflect the model predictions with and without the i-th feature, respectively.

Calculating exact SHAP values is computationally intensive. To approximate these values, Lundberg and Lee[73] and Lundberg *et al.*[78] introduced three primary methods: KernelSHAP, TreeSHAP, and DeepSHAP. KernelSHAP is a model-agnostic approach, utilizing a weighted linear regression to estimate Shapley values across models. TreeSHAP offers computational advantages specifically for tree-based models, while DeepSHAP, integrating DeepLIFT with the SHAP framework, is tailored for deep neural networks[73]. Therefore, in this study, given the diversity of models employed, from SVM to tree-based models, we implement the KernelSHAP method to maintain model compatibility. The SHAP values were computed using the KernelSHAP package in R 4.3.2[79].

Data Records

The dataset is available at Science Data Bank[80]. Our dataset offers a comprehensive overview spanning 70 years of China's policy stringency on climate change, from 1954 to 2022. The dataset comprises all policies with potential climate change implications promulgated by the central government, along with their corresponding policy stringency. The unique policy number in the dataset (for example, 201317437) signifies the yearly count of archived policies. The repository includes the following datasets:

- Measurement and Decomposition of China's PSCC (1954–2022): The PSCC data is measured and broken down into detailed sub-dimensions, facilitating a causal analysis of policy impacts and trends over the specified period and fields.
- Policy Pool with 2,216 Records: This dataset includes cleaned and organized 2,216 records, providing high-quality text data for other derivative analyses.
- Multi-layered, Cross-sectoral Lexicon utilized in PSCC measurement with overall version and splitted version.
- README file for data records usage and further analysis instructions.
- Optimized Chain-of-Thought (CoT) Prompts for Filtering and Categorizing System: These prompts are designed to enhance the accuracy and efficiency of the filtering and categorization processes, ensuring that the system can effectively handle complex policy data and those detailed technic files can be found in Supplementary Information document.

Our dataset is organized in a year-series format (e.g., China's Policy Stringency on Climate Change 1954), making it readily accessible for analysis. Additional notes provide context and guidance for understanding the data: the 'Category I' column delineates different policy types, with '1' representing CCI, '2' denoting MXI, and '3' signifying another category. The 'Category II' column maps these policies to current mainstream fields

Dataset Name	Dataset Scope	Year Coverage	Source
Environmental Policy Stringency Index	Country level, all sectors with six dimensions, addressing three types of environmental policies.	1990–2020	OECD (2016)[81]
China's Environmental Policy Intensity for 1978–2019	Country level, all sectors with fifteen dimensions, addressing three types of environmental policies.	1978–2019	Zhang *et al.*[52]
China's Low-Carbon Policy Intensity Dataset from National- to Prefecture-Level over 2007–2022	Country, provincial, and city levels, manufacturing sector with six dimensions, addressing three types of low-carbon policies.	2007–2022	Dong *et al.*[50]
Harmonizing existing climate change mitigation policy datasets with a hybrid machine learning approach	Cross-Country levels, climate change mitigation policy dataset.	1974–2022	Wu *et al.*[82]

Table 3. The Existing Dataset Overview.

in climate change, while 'Category II Mapping' serves as a unique identifier for these mappings. The 'Manual Rating' and 'Machine Learning' columns present the overall stringency scores derived through separate evaluation methodologies.

Specifically, regarding the sub-dimension stringency column content, the '10_climate_mitigation_energy_energysystems_m' column quantifies the stringency of policies within the energy systems aspect of climate mitigation. The number '10' indicates the identification number of the sub-lexicon. The dataset's structure adheres to the categorization outlined in the chapter 'Multi-layered, cross-sectoral lexicon build-up,' dividing mitigation, adaptation, implementation, and spatial difference measurement into distinct sub-dimensions. Within this framework, the stringency of policies has been evaluated across 80 sub-dimensions, including 50 for mitigation, 23 for adaptation, 3 for implementation, and 4 for spatial differences. Each of the mitigation and adaptation dimensions can be further subdivided into policy objectives and policy measures, providing a nuanced and detailed assessment of policy stringency.

The provided data records can be utilized both separately and integratively, offering versatile applications for researchers. The PSCC Measurement Results are now equipped to support deeper causal analyses, enabling exploration of governmental behavior and attention toward climate change. By decomposing these results into sub-dimensions, scientists can investigate specific fields to analyze fluctuations and variations within the PSCC framework. This method not only facilitates detailed investigations but also serves as a template for organizing large volumes of unstructured data for robust quantitative analysis. The meticulously cleaned and organized policy text data is primed for diverse studies, including case studies, qualitative analyses, and sentiment analyses. The overall lexicon clearly demonstrates how the PSCC framework aligns with the mainstream consistency of the IPCC guidelines. Meanwhile, the split lexicon can be paired with sub-dimension measurement results to conduct advanced analyses such as causal, trend, and field studies. For example, researchers can explore the impact of climate change on human migration or analyze government behaviors characterized by dimensions 2113 (Adaptation_Cross-sectoral_Planned Relocation And Resettlement). Furthermore, the dataset outlined in this record offers diverse opportunities for research and application across multiple dimensions. One potential direction lies in advancing methodologies. Employing ensemble machine learning models and advanced techniques such as BERT embeddings could significantly improve PSCC prediction accuracy. Collaborating with experts in climate science, environmental studies, and economics could further refine the lexicon development process, enabling it to better address rapid technological advancements and policy shifts. Exploring spatial variations in policy stringency is another promising avenue. Incorporating provincial or city-level data could provide a deeper understanding of the spatial heterogeneity of policies, enabling more localized and actionable climate strategies. Such granular insights would allow researchers to examine the diverse impacts of policies across different regions. Future research could also focus on investigating the impacts and correlations of PSCC. Empirical studies that explore the relationships between PSCC and variables such as green innovation, economic productivity, and temperature changes could yield valuable findings. Additionally, further categorization of the PSCC database to assess firm-level impacts, along with analyses of the complex interplay between policy implementation and other dimensions in the lexicon, could provide deeper insights into the dynamics of policy design and enforcement.

Technical Validation

In this section, we begin by validating the performance of our LLM-based filtering and classification system, followed by an external benchmark validation of our PSCC measurements. First, we evaluate various prompt configurations to identify the optimal setup — an instruction-based prompt combined with few-shot, chain-of-thought (COT), and augmentation techniques — that yields the highest performance for policy filtering and policy instrument classification. This selected prompt configuration is then applied to filter and classify the remaining policies and policy instruments. Next, we assess several machine learning algorithms, ultimately selecting Random Forest as the model for PSCC measurement. Additionally, we perform SHAP (SHapley Additive exPlanations) analysis to identify variables with the most significant contributions to the model's predictions. Finally, to ensure the robustness of our PSCC predictions, we validate our results by comparing them to established datasets from Zhang *et al.*[52], OECD (2016)[81], and Dong *et al.*[50].

LLM filtering and classifying system performance. In validating the performance of our Large Language Model (LLM) classification system, we initially trained and evaluated the model on a subset of 800 labeled policies from our extensive policy inventory of over 24,000 entries. The dataset was divided, with 80%

Policy Filtering				
Prompt_Type	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Zero-Shot	0.781	0.981	0.605	0.748
One-Shot	0.831	0.968	0.709	0.819
Few-Shot	0.806	0.982	0.651	0.783
Instruction + One-Shot	0.850	0.984	0.733	0.840
Instruction + Few-Shot	0.881	0.959	0.814	0.881
Instruction + One-Shot + CoT	0.888	0.972	0.814	0.886
Instruction + Few-Shot + CoT	0.888	0.947	0.837	0.889
Instruction + Few-Shot + CoT + Augmentation	0.931	1.000	0.872	0.932

Table 4. The LLM Policy Filtering Performance. Note: For policy filtering, binary precision, recall, and F1-score were used.

Policy Instrument Classification				
Prompt_Type	**Accuracy**	**Precision**	**Recall**	**F1_Score**
Zero-Shot	0.938	0.937	0.938	0.938
One-Shot	0.925	0.924	0.925	0.912
Few-Shot	0.925	0.917	0.925	0.917
Instruction + One-Shot	0.944	0.941	0.944	0.940
Instruction + Few-Shot	0.938	0.938	0.938	0.938
Instruction + One-Shot + CoT	0.944	0.941	0.944	0.940
Instruction + Few-Shot + CoT	0.944	0.943	0.944	0.943
Instruction + Few-Shot + CoT + Augmentation	0.956	0.951	0.956	0.950

Table 5. The LLM Policy Instrument Classification Performance. Note: For policy instrument classification, weighted-averaged metrics were applied.

allocated for training and the remaining 20% reserved for validation. To explore the model's capabilities, we tested seven different prompt scenarios, ranging from zero-shot learning to more sophisticated configurations such as instruction + few-shot + Chain-of-Thought (COT) + data augmentation. The evaluation metrics for filtering and classification performance were based on binary-average and weighted-average metrics, respectively, to account for any class imbalance within the dataset.

The results, summarized in Tables 4, 5, and Fig. 7, indicate that the instruction + few-shot + COT + augmentation prompt scenario consistently outperformed the others across all evaluation metrics for both policy classification and policy instrument classification tasks. To further substantiate these findings, we conducted an independent manual evaluation. We randomly selected 500 policies from the 24,000 + policies classified by the system, employing stratified random sampling to ensure a balanced representation. This sample comprised 250 policies predicted to have climate change mitigation or adaptation effects and 250 policies without such effects. For policy instrument classification, we selected an additional 500 policies from the 2,100 + policy pool, which included 441 command-and-control policies, 31 market-based policies, and 28 voluntary policies. These samples were manually labeled, and the manually labeled results were compared against the model's predictions.

Table 6 provides an accurate comparison between the LLM predictions and the manual labeling, highlighting the strong alignment between the two. This further confirms the robustness and reliability of our LLM-based classification system. In summary, the LLM classification system was rigorously validated through a combination of internal performance metrics and external manual checks, resulting in the successful classification of over 2,100 policies with potential climate change mitigation or adaptation effects. This includes 2,007 command-control policies (CCI), 152 market-based policies (MXI), and 57 voluntary policies (VOI).

Machine learning model performance.

The model we trained contains a range of algorithms, from simple linear regression to regularized approaches such as Lasso and Elastic Net, and extends to advanced machine learning algorithms like Support Vector Machines (SVM) and Random Forest (RF). The performance of these machine learning models is summarized in Fig. 4 and Table S9 (see Supplementary Information document). Notably, Random Forest demonstrated the lowest Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and residual compared to other models. Consequently, we opted to employ Random Forest for evaluating our PSCC.

SHAP analysis.

After showing that learning-based models can estimate the PSCC, we applied the kernal-SHAP to extract the important features in estimating PSCC using Random Forest. We summarised our SHAP value of the Random Forest model in Fig. 5.

From Plot (a) of Fig. 5, the value associated with the mitigation dimension shows a comparatively higher contribution to estimating the PSCC than the adaption and implementation dimension. Specifically, the word

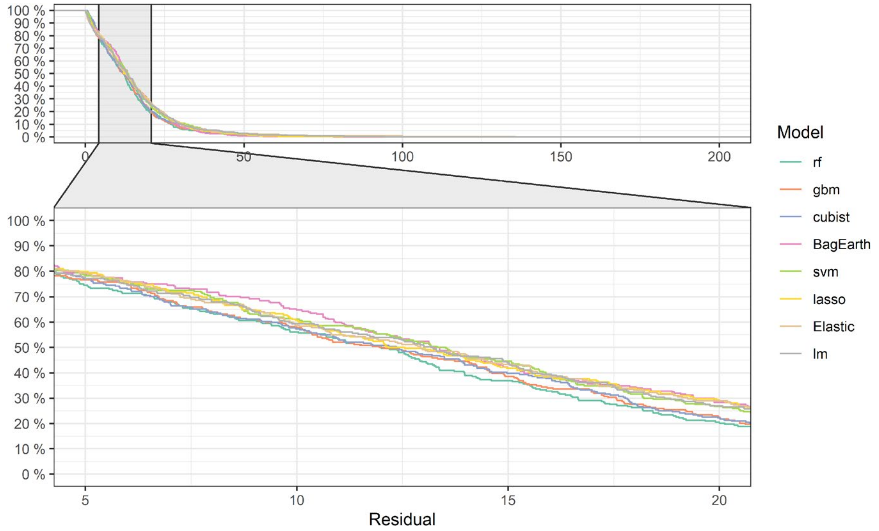


Fig. 4 The Model Performance. This diagram compares the residual distribution of various models: Random Forest (RF), eXtreme gradient boosting with tree booster (GBM), cubist regression model (Cubist), a bagging wrapper for multivariate adaptive regression splines via the earth function (BagEarth), Support Vector Machine (SVM), Lasso Regression (lasso), Elastic Net (Elastic), and Linear Model (LM). The x-axis shows residuals, and the y-axis displays the percentage of data with residuals at least as large. Steeper curves indicate fewer large residuals, signifying better model performance.

	Manual Checking for Classification	
	Policy Filtering	**Policy Instrument Classification**
	(1)	**(2)**
Accuracy	0.874	0.962
Precision	0.748	0.961
Recall	1.0	0.962
F1_Score	0.866	0.961

Table 6. Manual Checking for Filtering and Classification. Note: For policy filtering, binary precision, recall, and F1-score were used. For policy instrument classification, weighted-averaged metrics were applied.

related to energy (coal, energy systems, bio-electricity, and geothermal energy), AFOLU(reforestation), industry(recycling, industry system, and energy efficiency), spatial difference, and implementation (consistent and permissive) has a higher contribution to estimating the PSCC. Notably, from the policy measure and objective perspective, the contribution of the policy measure is higher than the policy objective.

We have illustrated the top 8 features in the SHAP dependence plot, as shown in Plot (b) of Fig. 5. In this plot, each dot represents the SHAP value of a variable, with the color indicating varying PSCC values. This visualization effectively demonstrates how the importance of each variable shifts with changes in both its own value and the PSCC value. For features such as industry system, energy system, energy coal, industry energy efficiency, industry recycling, afolu reforestation, and cross-sectoral climate services, higher values in these mitigation categories are associated with an increase in PSCC. Notably, the fluorinated feature exhibits a more scattered pattern in SHAP values, particularly at lower feature values, indicating inconsistent influence. Additionally, climate services show a mild positive relationship between feature values and SHAP values, with relatively low variability.

Comparison with existing datasets.

Due to the lack of a publicly accessible dataset that covers policy stringency on climate change (PSCC), direct comparisons with existing datasets are not feasible. However, there are some related datasets. Table 3 summarizes those datasets, detailing their scope and coverage years. Both OECD (2016) and Zhang *et al.*[52] explore the environmental policy intensity on the country level with different dimensions and policy categories. Dong *et al.*[50] developed a dataset on low-carbon policy intensity from 2007 to 2022 across the country, provincial, and city-level on manufacturing sector with different dimensions and policy categories. Wu *et al.*[82] developed the comprehensive climate policy dataset across different countries but without any intensity or stringency evaluation. Due to methodological and policy similarities with Zhang *et al.*[52], dimensional similarities with OECD (2016)[81], and topic similarities with Dong *et al.*[50], we selected those three datasets for comparison.

Following the approach of Dong *et al.*[50], we utilize Dynamic Time Warping (DTW) to measure the similarity between three series that exhibit varying trends. After normalizing three indices, we calculate three pairs of

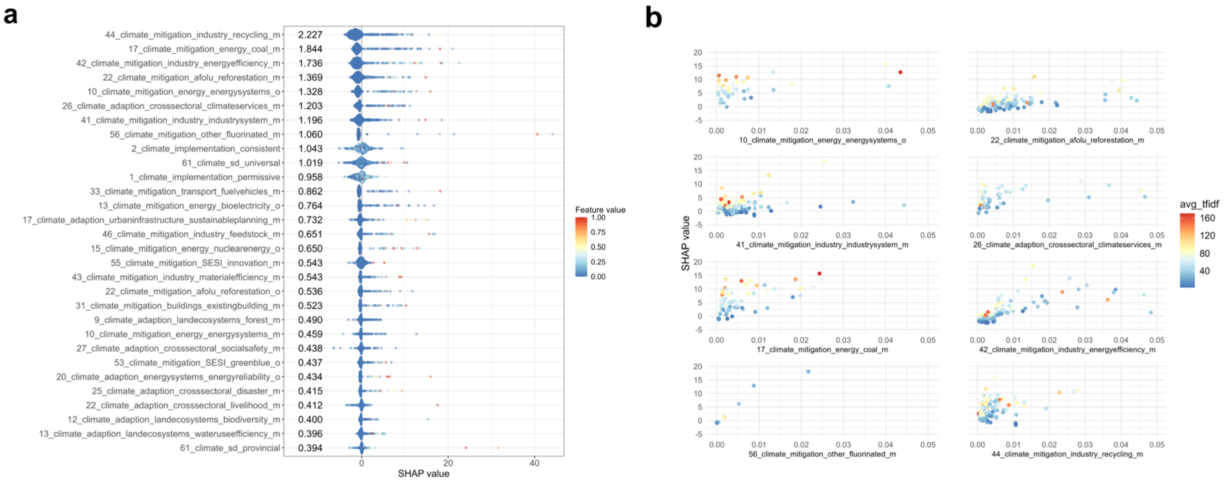


Fig. 5 SHAP Analysis of Feature Importance and Dependence. This composite figure combines a distribution plot and multiple dependence plots of SHAP values to illustrate the influence of PSCC features within an RF model. Plot (**a**) showcases the distribution of SHAP values for critical features, with the x-axis measuring the impact of these features on model output and the y-axis ranking the features by their importance. The color gradient from purple to yellow illustrates the range of feature values, aiding in the identification of features with the strongest predictive power. Plot (**b**) includes dependence plots for select high-importance features, depicting the relationship between SHAP values and feature values. The color scale indicates the feature's PSCC, thereby demonstrating how variations in feature values correspond to the model's output.

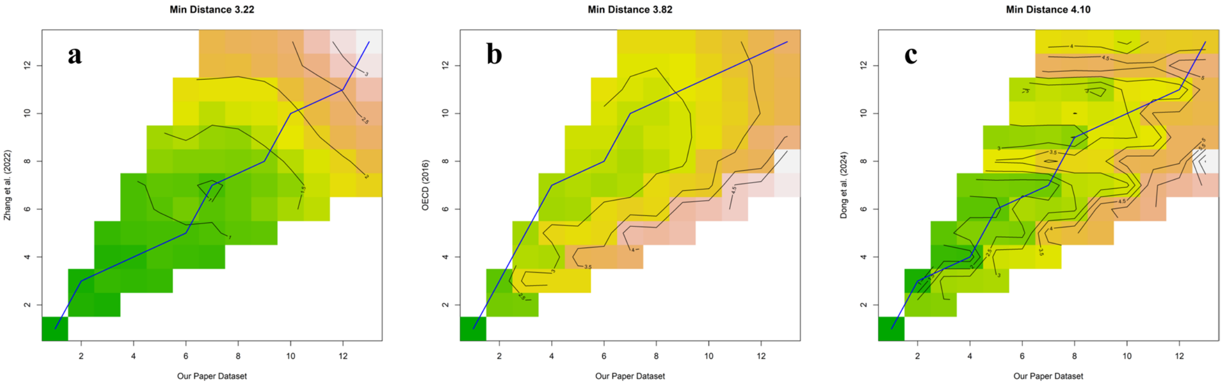


Fig. 6 Dynamic Time Wrapping Plot. The contour plots illustrate the comparison between the dataset used in our paper (horizontal axis) and datasets from three other sources (vertical axis). (**a**) Comparison Between Zhang *et al.*[52]: The contour plot shows the differences between our dataset and the one used by Zhang *et al.*[52]. The minimum distance observed is 3.22. (**b**) Comparison Between OECD (2016): The comparison between our dataset and the OECD (2016) dataset shows a minimum distance of 3.82. (**c**) Comparison Between Dong *et al.*[50]: The comparison between our dataset and Dong *et al.*[50] reveals the minimum distance of 4.10. The blue lines represent the path of minimal distance across different stringency levels, with contour levels highlighting areas of divergence and convergence.

DTW distances using the Euclidean distance. The paths that minimize the DTW distances between these indices are illustrated in Fig. 6. The distance ranges from 3.22 to 4.10 and features DTW minimum paths represented by a blue line that is small and nearly parallel to the main diagonal. This indicates a high degree of trend similarity between the national-level indices in this paper and the other three. This suggests that the PSCC developed in this paper effectively reflects the evolving trend of policy stringency.

We also show the overall trend of our datasets and three national indices. As depicted in Figure S3 (see Supplementary Information document), our PSCC and the other three datasets both show a general linear increase, and there are similarities and differences shared between our PSCC and the other two datasets.

Firstly, the EPS from Zhang *et al.*[52] fitted well with our PSCC. Significant climate events such as the establishment of the IPCC, the release of IPCC's first assessment report (AR1) and the formation of China's National Climate Change Coordination Group (NCCG), the inaugural Conference of the Parties (COP1), and the release of the fourth assessment report alongside the Bali Road Map at COP13, all correspond to similar peaks or troughs in both datasets. However, two notable divergences are observed. One is the United Nations Framework Convention on Climate Change entered into force (1993), another is the Marrakesh Accords, and the third

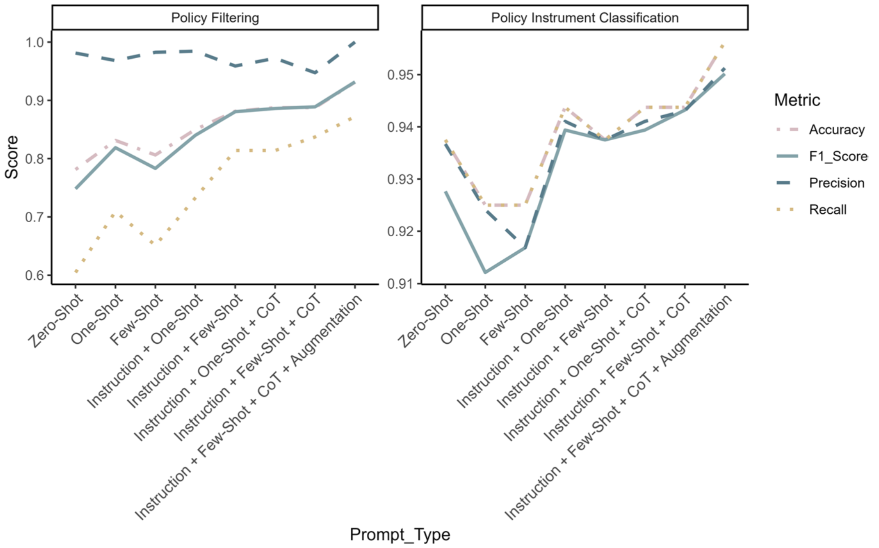


Fig. 7 LLM Accuracy for Policy Filtering and Policy Instrument Classification. The graph illustrates the accuracy in two tasks: policy filtering based on climate change mitigation/adaptation effects and policy instrument classification (CCI, MXI, VOI). The left-hand side of the graph represents the accuracy metrics using binary-average method for policy filtering, while the right-hand side displays the accuracy metrics using weighted-average method for policy instrument classification. The different metrics represented by the different types of line. For policy filtering, binary-averaged precision, recall, and F1-score were used. For policy instrument classification, weighted-averaged precision, recall, and F1-score were applied.

assessment report (2001), where EPS exhibits a trend contrary to PSCC. Both disparities stem from China's exclusive focus on environmental policies before 2001 and the prevailing ambiguity between environmental and climate policies until then.

Secondly, the EPS from OECD (2016)[81] did not align well with either our PSCC or the EPS from Zhang *et al.*[52]. This divergence primarily stems from the differing methodologies employed in evaluating policy stringency. The EPS by OECD (2016)[81] primarily utilizes statistical measures such as taxes, certificates, performance standards, R&D expenditure, etc. In contrast, both our study and that of Zhang *et al.*[52] adopt machine learning and text-analysis approaches for evaluating PSCC and EPS.

Thirdly, while Dong *et al.*[50] show larger fluctuations compared to our datasets, several key event-driven trends align our dataset with their dataset, such as the establishment of the Climate Change Department within China's NDRC in 2008, the release of the National Climate Change Response Plan (2014–2020) in 2014, and the announcement of China's carbon peak and neutrality goals for 2030 and 2060 in 2020. There is a discrepancy during the period from 2013 to 2020. The primary reason for the discrepancies is that, during the period from 2013 to 2017, China adopted a highly environmental focus, with policies heavily oriented toward combating air, water, and solid waste pollution. This orientation emphasized reforms within the manufacturing sector like strengthening industrial emission standards and upgrades on industrial boilers[83].

As a result, the intense environmental focus in China during this period significantly influenced Dong *et al.*'s findings, leading to discrepancies between their trends and those observed in our broader dataset. Consequently, a comparison between our PSCC and the EPS from OECD (2016)[81] underscores the advantages of policy stringency evaluation through text analysis. This method demonstrates a superior capability in capturing subtle shifts in policy topics and directions, which numerical measures may overlook. For instance, the spillover effects of the COVID-19 pandemic are evident in the trend of our PSCC dataset. Additionally, following China's announcement of its carbon peak and neutrality goals for 2030 and 2060, there is a noticeable and continuous uptrend in PSCC.

Acknowledging the similar methodologies employed in the dataset Zhang *et al.*[52], we utilized the 'fuzzyjoin' package developed by Robinson[84] in R 4.3.2 to identify the same policies in both datasets for a deeper examination of the discrepancies between PSCC and EPS in the same policy. The findings are depicted in Figure S4 (see Supplementary Information document) in the supplementary material. From this figure, a distinct divergence around the year 2000 becomes apparent. Before 2000, the discrepancy between PSCC and EPS was small in the same policy, but post-2000, we observed a clear bifurcation, indicating the formal separation of climate and environmental policies in China. And this discovery further explained the divergence we found in Figure S3 (see Supplementary Information document).

Code availability

Supplementary material to this article can be found on uploaded documents and https://doi.org/10.57760/sciencedb.14709.

Received: 10 June 2024; Accepted: 10 January 2025;
Published online: 31 January 2025

References

1. Stechemesser, A. *et al*. Climate policies that achieved major emission reductions: Global evidence from two decades. *Science* **385**(6711), 884–892 (2024).
2. Maamoun, N. The Kyoto Protocol: Empirical evidence of a hidden success. *Journal of Environmental Economics and Management* **95**, 227–256 (2019).
3. Eskander, S. M. & Fankhauser, S. Reduction in greenhouse gas emissions from national climate legislation. *Nature Climate Change* **10**(8), 750–756 (2020).
4. Dubash, N. *et al*. In: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 13. Cambridge University Press (2022).
5. Althammer, W. & Hille, E. Measuring climate policy stringency: A shadow price approach. *International Tax and Public Finance* **23**, 607–639 (2016).
6. Galeotti, M., Salini, S. & Verdolini, E. Measuring environmental policy stringency: Approaches, validity, and impact on environmental innovation and energy efficiency. *Energy Policy* **136**, 111052 (2020).
7. Botta, E. & Koźluk, T. Measuring environmental policy stringency in OECD countries: A composite index approach. OECD (2014).
8. You, X. AI analysed 1,500 policies to cut emissions. These ones worked. *Nature* (2024).
9. Bashmakov, I. *et al*. Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Chapter 11 (2022).
10. Pörtner, H.-O. *et al*. IPCC, 2022: Summary for policymakers. Cambridge University Press (2022).
11. Bento, A. M., Miller, N., Mookerjee, M. & Severnini, E. Incidental adaptation: The role of non-climate regulations. *Environmental and Resource Economics* **86**(3), 305–343 (2023).
12. Walker, B. J., Kurz, T. & Russel, D. Towards an understanding of when non-climate frames can generate public support for climate change policy. *Environment and Behavior* **50**(7), 781–806 (2018).
13. Wu, S. A systematic review of climate policies in China: Evolution, effectiveness, and challenges. *Environmental Impact Assessment Review* **99**, 107030 (2023).
14. Minx, J. C. *et al*. A comprehensive and synthetic dataset for global, regional, and national greenhouse gas emissions by sector 1970–2018 with an extension to 2019. *Earth System Science Data* **13**(11), 5213–5252 (2021).
15. Crippa, M., Solazzo, E., Guizzardi, D., Tubiello, F. N. & Leip, A. Climate goals require food systems emission inventories. *Nature Food* **3**(1), 1–1 (2022).
16. Zhang, W.-W., Xu, S.-C. & Sharp, B. Do CO2 emissions impact energy use? An assessment of China evidence from 1953 to 2017. *China Economic Review* **57**, 101340 (2019).
17. Gallagher, K. S., Zhang, F., Orvis, R., Rissman, J. & Liu, Q. Assessing the policy gaps for achieving China's climate targets in the Paris Agreement. *Nature Communications* **10**(1), 1256 (2019).
18. Teng, F. & Wang, P. The evolution of climate governance in China: Drivers, features, and effectiveness. *Environmental Politics* **30**(sup1), 141–161 (2021).
19. Wang, P., Liu, L. & Wu, T. A review of China's climate governance: State, market, and civil society. *Climate Policy* **18**(5), 664–679 (2018).
20. Wang, H. & Chen, W. Gaps between pre-2020 climate policies with NDC goals and long-term mitigation targets: Analyses on major regions. *Energy Procedia* **158**, 3664–3669 (2019).
21. Yu, B. Bureaucratic deliberation and China's engagement with international ideas: A case study on China's adoption of carbon emissions trading. *Journal of Contemporary China* **31**(136), 558–573 (2022).
22. Lee, H. *et al*. IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the IPCC. Geneva: Intergovernmental Panel on Climate Change (IPCC) (2023).
23. Friedlingstein, P. *et al*. Global carbon budget 2020. *Earth System Science Data Discussions* **2020**, 1–3 (2020).
24. Friedlingstein, P. *et al*. Global carbon budget 2021. *Earth System Science Data* **14**(4), 1917–2005 (2022).
25. Harrison, J., Kim, J. Y. & Völkle, J. RSelenium: R Bindings for 'Selenium WebDriver'. R package version 1.7.9 (2023).
26. Argyle, L. P. *et al*. Out of one, many: Using language models to simulate human samples. *Political Analysis* **31**(3), 337–351 (2023).
27. Xiaofei, S. *et al*. Text classification via large language models. arXiv preprint arXiv:2305.08377 (2023).
28. Gunes, E. & Florczak, C. K. Multiclass classification of policy documents with large language models. *arXiv preprint* arXiv:2310.08167 (2023).
29. Chae, Y. & Davidson, T. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation* (2023).
30. Liu, P. *et al*. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* **55**(9), 1–35 (2023).
31. Raffel, C. *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67 (2020).
32. Ainslie, J. *et al*. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint* arXiv:2305.13245 (2023).
33. Touvron, H. *et al*. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint* arXiv:2307.09288 (2023).
34. Achiam, J. *et al*. GPT-4 technical report. *arXiv preprint* arXiv:2303.08774 (2023).
35. Yang, A. *et al*. Qwen2 technical report. *arXiv preprint* arXiv:2407.10671 (2024).
36. Fu, Z. *et al*. Decoder-only or encoder-decoder? Interpreting language model as a regularized encoder-decoder. *arXiv preprint* arXiv:2304.04052 (2023).
37. Wei, J. *et al*. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837 (2022).
38. Howlett, M., Ramesh, M., & Perl, A. Studying public policy: Policy cycles and policy subsystems. Oxford University Press, Toronto (1995).
39. Housley, W. & Fitzgerald, R. Categorization, interaction, policy, and debate. *Critical Discourse Studies* **4**(2), 187–206 (2007).
40. Throgmorton, J. A. Reviews: Policy design for democracy. Anne Larason Schneider and Helen Ingram. Lawrence, KS: University Press of Kansas. 241 pages. *Journal of Planning Education and Research* **18**(3), 275–277 (1999).
41. Bemelmans-Videc, M.-L., Rist, R. C., & Vedung, E. The sermon: Information programs in the public policy process—choice, effects, and evaluation. In: *Carrots, Sticks, and Sermons*, pp. 103–128. Routledge (2017).
42. Smith, Z. A., & Jacques, P. The environmental policy paradox. Taylor & Francis (2022).
43. Boasson, E. L. & Tatham, M. Climate policy: From complexity to consensus? *Journal of European Public Policy* **30**(3), 401–424 (2023).
44. Hughes, L. & Urpelainen, J. Interests, institutions, and climate policy: Explaining the choice of policy instruments for the energy sector. *Environmental Science & Policy* **54**, 52–63 (2015).
45. Dong, Q. *et al*. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 1107–1128).
46. Wei, J. *et al*. Finetuned language models are zero-shot learners. *arXiv preprint* arXiv:2109.01652 (2021).
47. Chang, Y. *et al*. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* **15**(3), 1–45 (2024).

48. Rouge, L. C. A package for automatic evaluation of summaries. In: *Proceedings of Workshop on Text Summarization of ACL*, **vol. 5** (2004).
49. Jensen, J. L., Karell, D. & Tanigawa-Lau, C. Language models in sociological research: An application to classifying large administrative data and measuring religiosity. *Sociological Methodology* **52**(1), 30–52 (2022).
50. Dong, X., Wang, C., Zhang, F., Zhang, H. & Xia, C. China's low-carbon policy intensity dataset from national-to-prefecture-level over 2007–2022. *Scientific Data* **11**(1), 213 (2024).
51. Hossin, M. & Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* **5**(1), 1 (2015).
52. Zhang, G. *et al*. China's environmental policy intensity for 1978–2019. *Scientific Data* **9**(1), 123 (2022).
53. Zhang, G., Zhang, P., Zhang, Z. G. & Li, J. Impact of environmental regulations on industrial structure upgrading: An empirical study on Beijing-Tianjin-Hebei region in China. *Journal of Cleaner Production* **238**, 117848 (2019).
54. Likert, R. A technique for the measurement of attitudes. *Archives of Psychology* **22**(140), 1–55 (1932).
55. Tourangeau, R. The psychology of survey response. Cambridge University Press (2000).
56. Allen, I. E. & Seaman, C. A. Likert scales and data analyses. *Quality Progress* **40**(7), 64–65 (2007).
57. Dubash, N. K. *et al*. National and sub-national policies and institutions. In: *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the IPCC*. Cambridge University Press (2022).
58. Lim, B., Spanger-Siegfried, E., Burton, I., Malone, E., & Huq, S. Adaptation policy frameworks for climate change: Developing strategies, policies and measures. Cambridge University Press (2005).
59. Ara Begum, R. *et al*. 2022: Point of Departure and Key Concepts. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, chap. 1. Cambridge University Press (2022).
60. Averchenkova, A., & Chan, T. Governance pathways to credible implementation of net zero targets. Grantham Research Institute, London School of Economics (2023).
61. Mugambwa, J. *et al*. Policy implementation: Conceptual foundations, accumulated wisdom, and new directions. *Journal of Public Administration and Governance* **8**(3), 211–212 (2018).
62. Fransen, T. *et al*. Taking stock of the implementation gap in climate policy. *Nature Climate Change* **13**(8), 752–755 (2023).
63. Brysbaert, M., Mandera, P. & Keuleers, E. The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science* **27**(1), 45–50 (2018).
64. Baker, S. R., Bloom, N. & Davis, S. J. Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics* **131**(4), 1593–1636 (2016).
65. Lee, K., Choi, E. & Kim, M. Twitter-based Chinese economic policy uncertainty. *Finance Research Letters* **53**, 103627 (2023).
66. Lin, B. & Zhao, H. Tracking policy uncertainty under climate change. *Resources Policy* **83**, 103699 (2023).
67. Tao, T. E E pluribus unum: from complexity, universality. *Daedalus* **141**(3), 23–34 (2012).
68. Zhou, Z. *et al*. Quantitative analysis of Citi's ESG reporting: LDA and TF-IDF approaches. *Financial Engineering and Risk Management* **6**(3), 53–63 (2023).
69. Hartvig, Á. D., Pap, Á. & Pálos, P. EU Climate Change News Index: Forecasting EU ETS Prices with Online News. *Finance Research Letters* **54**, 103720 (2023).
70. Harrison, J. S., Thurgood, G. R., Boivie, S. & Pfarrer, M. D. Measuring CEO personality: Developing, validating, and testing a linguistic tool. *Strategic Management Journal* **40**(8), 1316–1330 (2019).
71. Athey, S. & Imbens, G. W. Machine learning methods that economists should know about. *Annual Review of Economics* **11**(1), 685–725 (2019).
72. Storm, H., Baylis, K. & Heckelei, T. Machine learning in agricultural and applied economics. *European Review of Agricultural Economics* **47**(3), 849–892 (2019).
73. Lundberg, S., & Lee, S.-I. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, **vol. 30** (2017).
74. Shapley, L. S. A value for n-person games. In: *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, pp. 31–40. Cambridge University Press (1988).
75. Masmoudi, S. *et al*. A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. *Science of the Total Environment* **715**, 136991 (2020).
76. Ramirez, S. G., Hales, R. C., Williams, G. P. & Jones, N. L. Extending SC-PDSI-PM with neural network regression using GLDAS data and Permutation Feature Importance. *Environmental Modelling & Software* **157**, 105475 (2022).
77. Gholami, H. *et al*. Interpretability of simple RNN and GRU deep learning models used to map land susceptibility to gully erosion. *Science of the Total Environment* **904**, 166960 (2023).
78. Lundberg, S. M. *et al*. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56–67 (2020).
79. Mayer, M., & Watson, D. Kernelshap: Kernel SHAP. R package version 0.3.8 (2023).
80. Fu, E. & Li, B. China's policy stringency on climate change. *Science Data Bank Repository*. https://doi.org/10.57760/sciencedb.14709 (2023).
81. OECD. Environmental policy: Environmental Policy Stringency Index. *OECD* (2016).
82. Wu, L., Huang, Z., Zhang, X. & Wang, Y. Harmonizing existing climate change mitigation policy datasets with a hybrid machine learning approach. *Scientific Data* **11**(1), 580 (2024).
83. Zhang, Q. *et al*. Drivers of improved PM2.5 air quality in China from 2013 to 2017. *Proceedings of the National Academy of Sciences* **116**, 24463–24469 (2019).
84. Robinson, D. Fuzzyjoin: Join tables together on inexact matching. R package version 0.1.6 (2020).

Acknowledgements

The authors are grateful for financial support from National Key R&D Program of China (Grant Number 2022YFF0802504 and 2021YFC1809000), National Natural Science Foundation of China (Grant Number 72140004) and the Special fund project for environmental protection of Guangdong province (2023-12, 2024-12).

Author contributions

Bo Li and Enxian Fu are co-first authors and contributed equally to this work. Wei Zhang and Jian Zhang are designated as co-corresponding authors, with Wei Zhang serving as the lead contact (zhangwei@caep.org.cn). Author contributions are as follows: (1) Bo Li & Enxian Fu: Methodology, Conceptualization, Analysis, Coding, Writing, and Editing. (2) Shuhao Yang, Jiaying Lin, Wei Zhang, Jian Zhang, Honqiang Jiang, Yaling Lu & Jiantong Wang: Supervision, Reviewing and Editing. The co-corresponding Jian Zhang made significant contributions during the revision process, particularly by providing the conceptual framework to address the spatial difference suggested by the reviewer. Additionally, Jian Zhang provided critical insights in supervising, reviewing and editing the manuscript. Based on these substantial contributions, the former co-corresponding

author, Hongqiang Jiang, has agreed to transfer his co-corresponding author designation to Jian Zhang. We also thank Na Li for offering valuable advice.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04476-0.

Correspondence and requests for materials should be addressed to W.Z. or J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.