Tracking Green Industrial Policies with LLMs: A Demonstration

Yucheng Lu
New York University, New York, USA
yuchenglu@nyu.edu

Abstract

Green industrial policies (GIPs) are government interventions that support environmentally sustainable economic growth through targeted incentives, regulations, and investments in clean technologies. As the backbone of climate mitigation and adaptation, GIPs deserve systematic documentation and analysis. However, two major hurdles impede this systematic documentation. First, unlike other climate policy documents, such as Nationally Determined Contributions (NDCs) which are centrally curated, GIPs are scattered across numerous government legislation and policy announcements. Second, extracting information from these diverse documents is expensive when relying on expert annotation. We address this gap by proposing *GreenSpyder*, an LLM-based workflow that monitors, classifies, and annotates GIPs from open-source information. As a demonstration, we benchmark LLM performance in classifying and annotating GIPs on a small expert-curated dataset. Our results show that LLMs can be quite effective for classification and coarse annotation tasks, though they still need improvement for more nuanced classification. Finally, as a real-world application, we apply *GreenSpyder* to U.S. Legislative Records from the 117th Congress, paving the way for more comprehensive LLM-based GIP documentation in the future. Code for this demonstration is publicly available at https://github.com/YuchengLu-NYU/GreenSpyderDemo.

1 Introduction

Climate change represents one of the most significant challenges of our time (Lee et al., 2023). Crucial to the mitigation and adaptation efforts are Green Industrial Policies (GIPs), which are "strategic government measures that aim to promote new economic sectors and accelerate structural change" towards a green economy (United Nations Environment Programme, 2024). GIPs encompass a wide range of governmental interventions, including targeted incentives, regulations, and investments in clean technologies. As economists and policy makers generally agree, these policies serve as the foundation for transitioning economies toward more sustainable practices while maintaining economic growth (Rodrik, 2014; Scoones et al., 2015; Ambec, 2017; Altenburg and Assmann, 2017). Despite their significance, there remains a substantial gap in the systematic documentation and analysis of GIPs. Current research predominantly examines isolated instances of GIPs rather than providing comparative analyses. For example, Partnership for Action on Green Economy (2019); Zeng et al. (2021) studied eco-industrial parks in China, while Choi and Qi (2019) studied the effectiveness of carbon trading in South Korea. A comprehensive cross-jurisdictional and temporal analysis would undoubtedly contribute to the formulation of evidence-based best practices and policy recommendations.

Unlike other climate policy instruments such as Nationally Determined Contributions (NDCs), which are centrally documented through international frameworks like the Paris Agreement (United Nations, 2015), GIPs lack a centralized repository. Instead, they are dispersed across various government publications, legislative records, and policy announcements, making comprehensive analysis challenging. Furthermore, the technical and domain-specific nature of these documents requires specialized knowledge to properly identify and categorize relevant policies, traditionally necessitating expensive expert annotation. To address these challenges, we propose *GreenSpyder*, a Large Language Model (LLM)-based workflow designed to monitor, classify, and annotate GIPs from open-source information. Our approach leverages recent advances in natural language processing (NLP) to automate much of the labor-intensive work of policy identification and classification, potentially en-

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 1–10
July 31, 2025 ©2025 Association for Computational Linguistics

abling more comprehensive and timely analysis of GIPs worldwide.

In this paper, we first evaluate the capability of LLMs in classifying and annotating GIPs using New Industrial Policy Observatory (NIPO), a small expert-curated dataset on industrial policies (Evenett et al., 2024). Our evaluation reveals that while LLMs perform well on differentiating GIPs from general industrial policies, and coarse annotation tasks, they still face limitations when handling more nuanced policy distinctions. Building on these insights, we demonstrate a practical application of our approach by applying *GreenSpyder* to U.S. Legislative Records from the 117th Congress, successfully identifying and annotating GIPs within this substantial corpus of legislative text.

Our work contributes to the growing intersection of NLP and climate policy (Stammbach et al., 2024; Singh et al., 2024; Joe et al., 2024; Gargiliotti, 2024) by providing a scalable method for GIP documentation, potentially enabling researchers, policymakers, and advocates to better track, compare, and analyze green industrial policies across different contexts. This improved visibility could ultimately support more effective policy design and implementation in the global effort to address climate change.

2 Methods

2.1 Workflow

Figure 1 illustrates the workflow of *GreenSpyder*. In the first step, *GreenSpyder* periodically scans and indexes new content from a source repository, which contains a list of expert-curated base URLs where information relevant to GIPs may be found. These sources include `https://govtrack.us` (U.S. Congressional Records), `https://ndrc.gov.cn` (China's National Development and Reform Commission), `https://commission.europa.eu` (European Commission), etc.

Subsequently, we leverage LLMs to filter GIP-relevant information and annotate key features for database storage. Light green nodes in the flowchart indicate components where LLMs may be integrated in future iterations. For instance, recent work by Lorenzo Padoan (2024) and UncleCode (2024) demonstrates LLM-powered scrapers that could enhance scraping and parsing accuracy. Similarly, during pre-processing, LLMs could facilitate translation into English before entering the processing pipeline, addressing the documented performance disparities between high-resource and low-resource languages in multilingual LLMs (Huang et al., 2023).

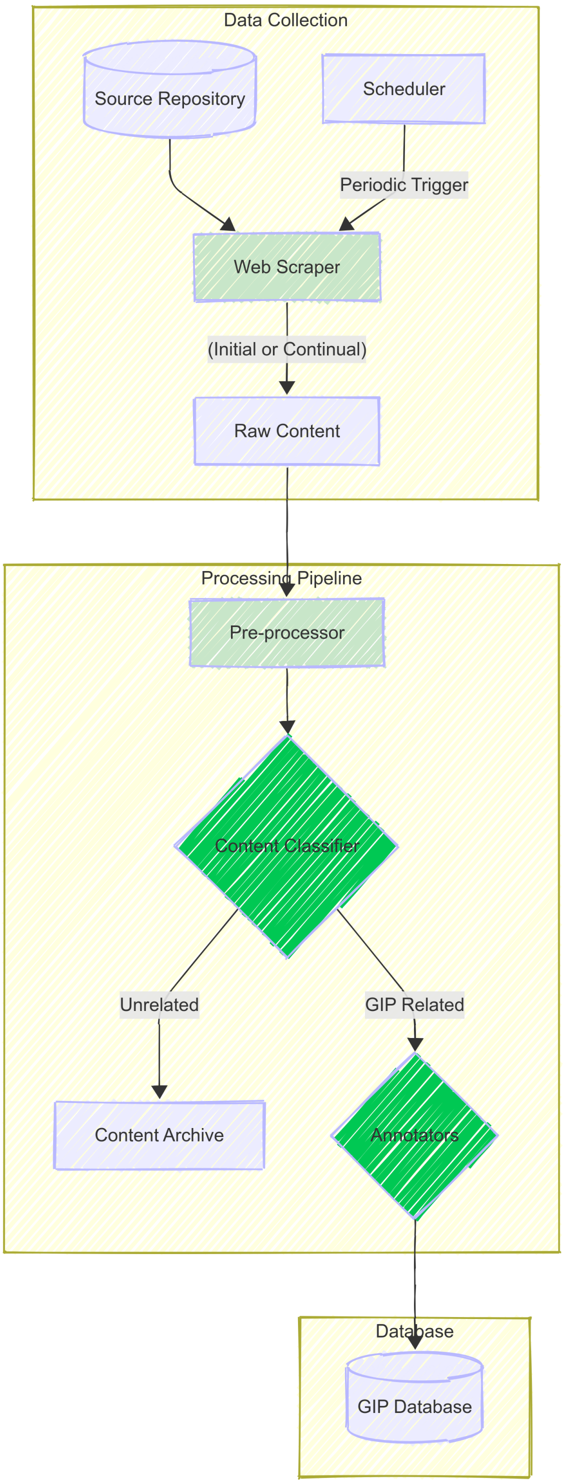


Figure 1: *GreenSpyder* Workflow

2.2 Experiments

Dark green nodes represent components where LLMs are currently implemented and constitute the focus of this demonstration. Specifically, we evaluate GPT-4o, a state-of-the-art LLM, as a few-shot classifier for identifying and annotating GIPs in one main task and three supplementary tasks, with increasingly complex analytical dimensions:

Main Task

- **Green Industrial Policy Classification (GIP)**: This foundational task requires the LLM to perform binary classification, distinguishing policy documents that constitute GIPs from those that do not. While seemingly straightforward, its accuracy is crucial as it serves as the initial filter in the GIP processing pipeline.

Supplementary Tasks

- **Targeted Jurisdiction Annotation (TJA)**: The LLM must identify specific jurisdictions (e.g., "European Union", "United States of America") targeted by a GIP. If no explicit jurisdiction is mentioned, the target is assumed to be the "Rest of the World" (ROW). On one hand, the fact that a single GIP can target multiple jurisdictions makes this a multi-label classification task, hence potentially challenging. On the other hand, however, the overall difficulty is expected to be medium to low, as it primarily leverages the LLM's general knowledge for recognizing named entities (countries, regions), requiring limited domain expertise in most instances.

- **Policy Instrument Annotation (PIA)**: This task involves categorizing GIPs into nine predefined policy instrument types (Export Policy, Import Policy, Trade Defense, Subsidy, Export Incentive, Procurement Policy, FDI Policy, Localization Policy, Other Policy). Detailed definitions of these instrument types are provided in Appendix B and are given to the LLM as part of the prompt. Widely used by economists (Criscuolo et al., 2022), this detailed taxonomy is crucial for analyzing the heterogeneous effects of different industrial policies and informing policy discussions. The primary challenge is interpreting policy language, which often uses euphemisms or technical jargon instead of explicit instrument labels. While structured as a multi-label classification (a policy could employ multiple instruments), in practice, many GIPs utilize a single primary instrument, making it often behave closer to a multi-class problem. Overall, we anticipate this to be a medium difficulty task for the LLM.

- **Harmonized System Annotation (HSA)**: The LLM is tasked with identifying specific products affected by GIPs, mapping them to the 6-digit Harmonized System (HS) code level. HS codes are internationally agreed product specifications and serve as a fundamental unit for economic analysis. This task tests the LLM's ability to bridge the gap between domain-specific policy terminology and the standardized international trade classification system. With over 5,000 product categories at the 6-digit level, this constitutes a demanding knowledge retrieval and mapping challenge, even for human experts. A significant constraint is that detailed descriptions of all HS codes cannot be provided to the LLM in-context due to prompt length limitations. We expect this to be a very challenging task via simple in-context learning.

We perform our experiments using the New Industrial Policy Observatory (NIPO) dataset.[1] NIPO is an expert-curated dataset that tracks industrial policies, created by the Global Trade Alert in collaboration with the International Monetary Fund. Crucially for our research, NIPO contains expert annotations that identify whether a policy qualifies as a Green Industrial Policy, the target jurisdictions, the type of policy instrument employed, and the impacted HS product codes. In total, the dataset contains 2,580 industrial policies, of which 439 are classified as GIPs.

Baseline Comparison For the main classification task, we finetune a RoBERTa-large model (Liu et al., 2019) using standard hyperparameters. To

[1] A publicly available subsample of the data can be found at `https://globaltradealert.org/reports/new-industrial-policy-observatory-nipo`. While GTA has tracked policy changes affecting global trade and investment since 2009, NIPO, which specifically focuses on industrial policies, only began in 2023. Moreover, since GTA's primary focus is on global trade and investment, they exclude policies that do not affect foreign interests, which means it does not provide a comprehensive database of GIPs but rather a select subsample.

Task	Classification Type	Domain Expertise	Label Space Size	Overall Difficulty
GIP	Binary	Low	Small	Low
TJA	Multi-label	Low	Medium	Low
PIA	Multi-label	Medium	Small	Medium
HSA	Multi-label	High	Large	High

Table 1: Comparison of expected task difficulties across classification type, required domain expertise, label space size, and overall difficulty.

mitigate small-sample issues, we apply Easy Data Augmentation (EDA) techniques from Wei and Zou (2019). Details about the finetuning procedure can be found in Appendix A.

However, for the supplementary tasks, finetuning RoBERTa proved impractical due to the limited size of the annotated dataset and the multi-label nature of these classification tasks. Instead, we offer a qualitative comparison of their expected difficulties, which are summarized in Table 1. This summary is based on an assessment of key task characteristics (classification type, required domain expertise, and label space size) and a heuristic estimation of manual annotation cost for each task, informed by our inspection of task requirements and some example policy texts.

Evaluation Metrics We use accuracy, macro-averaged F1 score, and hamming loss as our evaluation metrics. Hamming loss is specific to multi-label classification. It measures the fraction of incorrectly predicted labels in a multi-label classification task. It calculates the symmetric difference between predicted and true label sets, divided by the total number of labels. Formally, it is the proportion of labels that are incorrectly predicted (false positives and false negatives). Hamming loss ranges from 0 to 1, where 0 indicates perfect prediction and 1 indicates completely incorrect predictions. This metric is particularly suitable for multi-label tasks as it accounts for both missing relevant labels and incorrectly including irrelevant ones.

2.3 Application of *GreenSpyder*

Last but not the least, as a real-world application, we apply *GreenSpyder* to U.S. Legislative Records from the 117th Congress. 365 final bills (after consolidation and incorporation) were enacted during the 117th Congress. We scraped the content of these bills from `https://www.govtrack.us`. The goal is to identify and annotate GIPs from these 365 enacted bills.

3 Results

Table 2 illustrates the LLM's performance on the main task. GPT-4o achieved strong performance on the binary task of identifying Green Industrial Policies, with an accuracy of 0.94 and an F1 score of 0.90. This, in fact, slightly outperformed our finetuned RoBERTa-large baseline model, which potentially suffered from a lack of training data. The high performance on this foundational task establishes a reliable first stage in our processing pipeline.

Method	Accuracy	Macro F1
RoBERTa	0.92	0.89
GPT-4o	0.94	0.90

Table 2: Performance comparison on the Green Industrial Policy classification task. RoBERTa refers to a finetuned RoBERTa-large model, while GPT-4o results were obtained via few-shot prompting.

However, performance declines substantially for more complex annotation tasks requiring specialized domain knowledge, as Table 3 suggests.

Surprisingly, Target Jurisdiction Annotation (TJA) proved more challenging than initially anticipated, particularly when compared to Policy Instrument Annotation (PIA). For TJA, GPT-4o achieved an accuracy of only 0.31, a macro F1 score of 0.42, and a hamming loss of 0.42. These metrics collectively indicate significant difficulty: while the model might partially identify correct jurisdictions, it struggles to precisely capture all targeted regions. Several factors might contribute to this underperformance. These include potential mismatches in country naming conventions between the policy text and the ground truth labels; ambiguities in defining the precise target jurisdiction, such as when a supranational entity like the

EU provides a subsidy to companies within a member state; and inconsistencies in applying the "Rest of the World" (ROW) designation.

In contrast to TJA, for Policy Instrument Annotation (PIA), GPT-4o demonstrated more promising, albeit still intermediate, performance. The comparatively low hamming loss, in particular, indicates that even when the model does not identify all applicable policy instruments, its predictions are often reasonably close to the expert annotations. These results suggest a reasonable capability to interpret policy language and categorize interventions across the nine predefined instrument types despite the need for some domain expertise.

The most challenging task by far remained Harmonized System Annotation (HSA). Here, GPT-4o's performance dropped dramatically, achieving an accuracy of only 0.11, a macro F1 score of 0.12, and a high Hamming Loss of 0.69. This significantly lower performance compared to other tasks is largely attributable to the granularity of the HS taxonomy, which contains over 5,000 distinct product categories at the 6-digit level. However, to be fair to LLMs, HS code classification is also difficult for humans. Untrained individuals struggle significantly with this task, and even experts require reference materials to achieve accuracy.

Task	Accuracy	Macro F1	Hamming
TJA	0.31	0.42	0.42
PIA	0.65	0.67	0.32
HSA	0.11	0.12	0.69

Table 3: Performance on supplementary tasks. TJA: Target Jurisdiction Annotation. PIA: Policy Instrument Annotation. HSA: Harmonized System (product code) Annotation.

Application *GreenSpyder* identifies 6 GIPs from the 117th Congress, which are:

- H.R. 2471: Consolidated Appropriations Act
- H.R. 5376: Inflation Reduction Act
- H.R. 4346: CHIPS and Science Act
- H.R. 3684: Infrastructure Investment and Jobs Act
- S. 1605: National Defense Authorization Act
- H.R. 7776: James M. Inhofe National Defense Authorization Act

Upon manual inspection by the authors, all six identified bills were confirmed to contain provisions that align with the definition of GIPs. Notably, this set includes landmark legislation such as the Inflation Reduction Act and the CHIPS and Science Act, which are widely recognized for their significant GIP components, but also more obscure appropriations bills that contain GIP clauses (e.g., S. 1605: National Defense Authorization Act).

To further assess the classifier's specificity and guard against simply identifying any bill with environmental mentions, we conducted a qualitative analysis of potential false positives. We manually selected bills that contained keywords like "environment," "climate," or "energy" but were not classified as GIPs by *GreenSpyder*. For example:

- S. 1466 (Saline Lake Ecosystems in the Great Basin States Program Act) was correctly excluded. While environmentally focused, it primarily establishes a monitoring and assessment program rather than promoting specific green industries or technologies through industrial policy mechanisms.
- H.R. 1319 (American Rescue Plan Act of 2021) was also correctly excluded. While a major economic intervention (an industrial policy in a broad sense), its primary focus was on COVID-19 relief and economic recovery, lacking the specific green transition elements core to GIPs.

This initial check suggests that the system can differentiate GIPs from broader environmental legislation or general industrial policies that lack a green focus, indicating a degree of precision.

4 Conclusion

In this paper, we introduced *GreenSpyder*, an LLM-based workflow designed to systematically monitor, classify, and annotate Green Industrial Policies from diverse government sources. Our evaluation of GPT-4o on the expert-curated NIPO dataset demonstrated promising capabilities in distinguishing GIPs from general industrial policies and performing coarse-grained annotations, though challenges remain for more nuanced classification tasks. By successfully applying *GreenSpyder* to U.S. Legislative Records from the 117th Congress, we have demonstrated its practical utility in identifying and categorizing GIPs within large legislative corpora, offering a foundation for future advancements in automated GIP tracking.

5 Limitations

Despite the promising performance of *GreenSpyder* on the main GIP classification task, several limitations warrant careful consideration.

First, we did not apply the supplementary annotation tasks (TJA, PIA, HSA) to the U.S. Congressional Acts in our application. This was due in part to limited performance observed on these tasks in the NIPO dataset, and also because individual bills often bundle multiple interventions. Decomposing them into distinct GIP instances is a non-trivial challenge that our current workflow does not yet address. For example, a comprehensive piece of legislation like the U.S. Inflation Reduction Act contains numerous distinct provisions—such as tax credits for electric vehicle purchases, investments in renewable energy manufacturing, and funding for climate-smart agriculture—each potentially constituting a separate GIP with unique targets, instruments, and affected sectors, requiring a more granular level of analysis than simple bill-level classification.

Second, our evaluation relied on a relatively small, though expert-curated, dataset (NIPO). While useful for benchmarking, the dataset may underrepresent non-Western policy formats, informal legislation, or policies not tied to trade-impacting measures. This limits the generalizability of our findings to other jurisdictions or policy types.

Third, the "black-box" nature of large language models, particularly commercial ones like GPT-4o, complicates interpretability and debugging. As observed in our experiments, understanding failure modes—such as the underperformance of TJA relative to PIA—is difficult, limiting our ability to ensure consistent performance across domains.

These limitations point to key areas for future work, including scaling to multilingual or regionally diverse datasets, developing decomposition strategies for bundled legislation, and improving performance in granular annotation tasks.

6 Ethics

Closely related to the limitations discussed above, several ethical considerations arise in the development and potential deployment of *GreenSpyder*.

First, large language models may reflect and amplify existing global imbalances in data coverage. Since our demonstration relies on English-language sources and a dataset focused on internationally visible GIPs, the resulting annotations may overrepresent high-income, well-documented jurisdictions. This risks obscuring policy efforts from low-resource or non-English-speaking regions, thereby reinforcing unequal visibility in climate policy discourse.

Second, the use of automated policy monitoring tools, including web scraping, raises concerns about privacy and data sovereignty. While we restrict scraping to publicly accessible sources, care must be taken to avoid unintended surveillance or misuse of draft or sensitive policy documents that governments may be developing. Adherence to legal norms (e.g., `robots.txt`), institutional permissions, and ethical data sourcing practices is essential.

Third, automated classification tools can misinterpret or oversimplify complex policy language. If such outputs are used uncritically, they may influence downstream research or policy conclusions. To mitigate this, we emphasize that *GreenSpyder* is a research demonstration—not a production-ready tool or substitute for expert judgment. Human validation remains essential, particularly in high-stakes or ambiguous cases.

As LLMs continue to evolve, ongoing ethical review and engagement with a diverse range of stakeholders will be critical to ensuring responsible and equitable use in global policy analysis.

7 Acknowledgments

This research was supported in part by credits from OpenAI's Researcher Access Program. We also gratefully acknowledge the NYU IT High Performance Computing (HPC) team for providing computational resources, services, and technical expertise that facilitated this work. We thank the Global Trade Alert and the International Monetary Fund for curating the New Industrial Policy Observatory. Last but not least, we thank the anonymous reviewers for their helpful feedback.

References

Tilman Altenburg and Claudia Assmann. 2017. Green industrial policy. concept, policies, country experiences. Technical report, UN Environment; German Development Institute (DIE), Geneva, Bonn.

Stefan Ambec. 2017. Gaining competitive advantage with green industrial policy. In Tilman Altenburg and Claudia Assmann, editors, *Green Industrial Policy. Concept, Policies, Country Experiences*, pages 38–49. UN Environment; German Development Institute (DIE), Geneva, Bonn.

Y. Choi and C. Qi. 2019. Is south korea's emission trading scheme effective? an analysis based on the marginal abatement cost of coal-fueled power plants. *Sustainability*, 11(9):2504.

Chiara Criscuolo, Nicolas Gonne, Kohei Kitazawa, and Guy Lalanne. 2022. Are industrial policy instruments effective?: A review of the evidence in OECD countries. OECD Science, Technology and Industry Policy Papers 128, OECD Publishing.

Simon Evenett, Adam Jakubik, Fernando Martín, and Michele Ruta. 2024. The return of industrial policy in data. Working Paper 001, International Monetary Fund.

Dario Garigliotti. 2024. SDG target detection in environmental reports using retrieval-augmented generation with LLMs. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 241–250, Bangkok, Thailand. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Elphin Joe, Sai Koneru, and Christine Kirchhoff. 2024. Assessing the effectiveness of GPT-4o in climate change evidence synthesis and systematic assessments: Preliminary insights. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 251–257, Bangkok, Thailand. Association for Computational Linguistics.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barret, et al. 2023. Climate change 2023: Synthesis report, summary for policymakers. Ipcc report, Intergovernmental Panel on Climate Change (IPCC), Geneva, Switzerland.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marco Vinciguerra Lorenzo Padoan. 2024. Scrapegraph-ai. A Python library for scraping leveraging large language models.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Partnership for Action on Green Economy. 2019. Green transformation of industrial parks in jiangsu province: A synthesis report. Report, PAGE.

Dani Rodrik. 2014. Green industrial policy. *Oxford Review of Economic Policy*, 30(3):469–491. Accessed 31 Mar. 2025.

I. Scoones, M. Leach, and P. Newell, editors. 2015. *The Politics of Green Transformations*, 1 edition. Routledge.

Prashant Singh, Erik Lehmann, and Mark Tyrrell. 2024. Climate policy transformer: Utilizing NLP to track the coherence of climate policy documents in the context of the Paris agreement. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 1–11, Bangkok, Thailand. Association for Computational Linguistics.

Dominik Stammbach, Jingwei Ni, Tobias Schimanski, Kalyan Dutia, Alok Singh, Julia Bingler, Christophe Christiaen, Neetu Kushwaha, Veruska Muccione, Saeid A. Vaghefi, and Markus Leippold, editors. 2024. *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Association for Computational Linguistics, Bangkok, Thailand.

UncleCode. 2024. Crawl4ai: Open-source llm friendly web crawler & scraper. `https://github.com/unclecode/crawl4ai`.

United Nations. 2015. Paris agreement to the united nations framework convention on climate change.

United Nations Environment Programme. 2024. Green industrial policy.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Douglas Zhihua Zeng, Lei Cheng, Lei Shi, and Wilfried Luetkenhorst. 2021. China's green transformation through eco-industrial parks. *World Development*, 140:105249.

A Finetuning Details

For the GIP classification task, we finetuned a RoBERTa-large model (Liu et al., 2019). The dataset was split into training (80%) and validation (20%) sets. To address the limited size of the training data and improve generalization, we employed Easy Data Augmentation (EDA) techniques as proposed by Wei and Zou (2019). Specifically, we used EDA operations (Synonym Replacement, Random Insertion, Random Swap, and Random Deletion) with $\alpha = 0.05$ (the proportion of words altered per augmentation operation), and `num_aug=4`, generating four augmented versions for each original training sample.

The RoBERTa-large model was augmented with a linear classification head. The output representation of the `[CLS]` token was fed into this head, which includes a dropout layer with a ratio of 0.1 before the final classification layer. As is standard, we truncate input policy text to the first 512 tokens. As illustrated in Figure 2, the majority of policy texts in our dataset fall comfortably within this limit, minimizing information loss due to truncation. The model was trained for 3 epochs. We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-5}, a batch size of 16, and a weight decay of 0.01. A linear learning rate scheduler with a warm-up phase (10% of total training steps) was also employed.

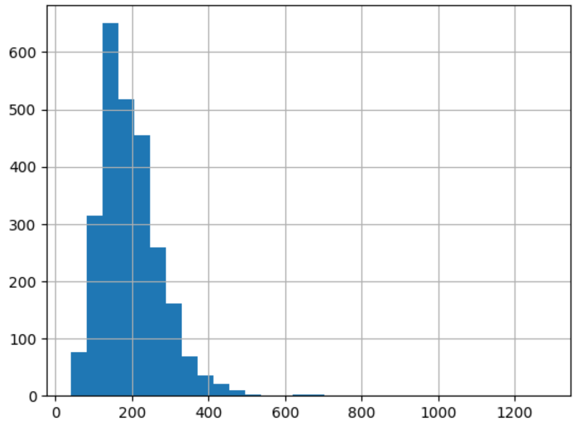
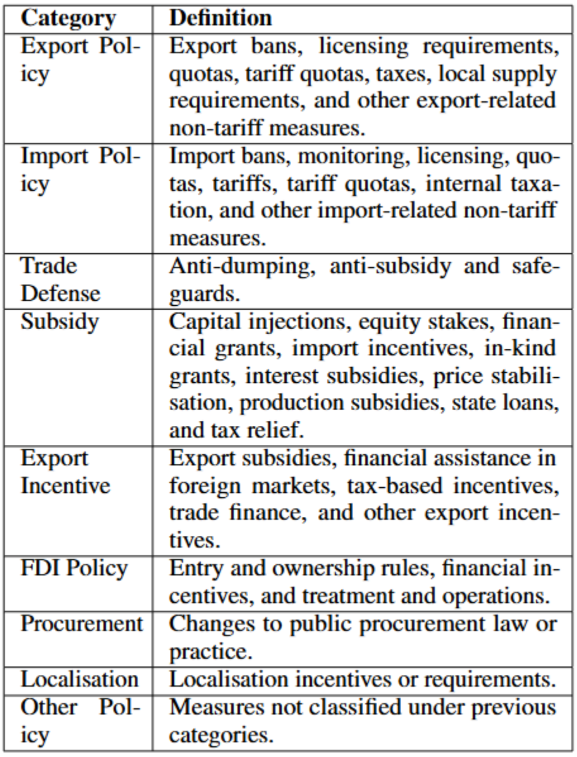


Figure 2: Histogram of Policy Text Length

B Additional Information about Policy Instrument Taxonomy



Category	Definition
Export Policy	Export bans, licensing requirements, quotas, tariff quotas, taxes, local supply requirements, and other export-related non-tariff measures.
Import Policy	Import bans, monitoring, licensing, quotas, tariffs, tariff quotas, internal taxation, and other import-related non-tariff measures.
Trade Defense	Anti-dumping, anti-subsidy and safeguards.
Subsidy	Capital injections, equity stakes, financial grants, import incentives, in-kind grants, interest subsidies, price stabilisation, production subsidies, state loans, and tax relief.
Export Incentive	Export subsidies, financial assistance in foreign markets, tax-based incentives, trade finance, and other export incentives.
FDI Policy	Entry and ownership rules, financial incentives, and treatment and operations.
Procurement	Changes to public procurement law or practice.
Localisation	Localisation incentives or requirements.
Other Policy	Measures not classified under previous categories.

Figure 3: Trade Policy Categories and Definitions. Source: New Industrial Policy Observatory (Evenett et al., 2024)

C Prompt Details

We use a few-shot prompting format for all tasks, where each input prompt contains three randomly sampled examples. Each example consists of a policy text excerpt and the corresponding expert-labeled response, placed directly before the test document. We randomize the examples for each inference call to reduce overfitting to specific prompts, though all are drawn from the training split of the NIPO dataset.

To ensure consistent and stable outputs, we set the generation temperature to 0.1 for all GPT-4o runs. This low temperature minimizes output variance and improves reproducibility, particularly important for classification and structured annotation tasks.

GIP Classification

You are an expert in industrial and environmental policy analysis. Your task is to determine whether the policy document provided below contains a Green Industrial Policy (GIP).

A Green Industrial Policy (GIP) is defined as:

- A government intervention aimed at promoting environmental sustainability while supporting industrial development
- Must have an explicit environmental focus (e.g., reducing emissions, promoting clean energy, improving resource efficiency)
- Must involve active industrial policy measures (subsidies, regulations, public investments, etc.)

Based on this definition, analyze the following policy document and determine whether it constitutes a GIP. Respond with "YES" if it is a GIP or "NO" if it is not.

Policy document: [POLICY TEXT]

Target Jurisdiction

You are an expert in international trade and industrial policy analysis. Your task is to identify all target jurisdictions specified in a Green Industrial Policy document.

Instructions:

- Read the policy document carefully
- Identify all jurisdictions (countries, regions, economic blocs) that are explicitly mentioned as targets of the policy.
- Write country names in their most common formats.
- If no specific jurisdictions are mentioned, assume the target is Rest of World (ROW)
- List all identified target jurisdictions, separated by commas
- If you identify ROW, list only ROW
- The target jurisdiction is defined as the geographical entity whose companies or industries are directly affected by the policy measures.

Policy document: [POLICY TEXT]

HS CODE

You are an expert in international trade classification systems, particularly the Harmonized System (HS) for product classification. Your task is to identify all 6-digit HS codes for products affected by a Green Industrial Policy document.

Instructions:

- Read the policy document carefully
- Identify all products or product categories mentioned in the document
- Determine the corresponding 6-digit HS codes for each identified product
- List all applicable 6-digit HS codes, separated by commas
- Use 2012 Harmonized System for product classification

Remember that HS codes follow a hierarchical structure:

- First 2 digits: Chapter (broad category)
- Digits 3-4: Heading (more specific category)
- Digits 5-6: Subheading (specific product)

Policy document: [POLICY TEXT]

Policy Instruments

You are an expert in industrial policy analysis. Your task is to classify a Green Industrial Policy document according to the types of policy instruments it employs.

A policy may employ multiple instruments. Please identify ALL that apply from the following categories:

- Export Policy: Measures affecting export operations (e.g., export taxes, restrictions, bans)
- Import Policy: Measures affecting import operations (e.g., tariffs, quotas, licensing requirements)
- Trade Defense: Measures to protect domestic industries from foreign competition (e.g., anti-dumping duties, countervailing measures)
- Subsidy: Direct financial support to companies or sectors (e.g., grants, loans, tax benefits)
- Export Incentive: Measures to promote exports (e.g., export credits, export guarantees)
- Procurement Policy: Government purchasing preferences or requirements
- FDI Policy: Measures affecting foreign direct investment (e.g., equity caps, local content requirements)
- Localization Policy: Measures requiring or encouraging local production or sourcing
- Other Policy: Any relevant policy instrument not covered above

List all applicable policy instruments, separated by commas.

Policy document: [POLICY TEXT]